

VCE UNITS 3 & 4

data analytics

5E

GARY BASS

NATALIE HEATH

THERESE KEANE

ANTHONY SULLIVAN

NELSON
A Cengage Company

Data Analytics VCE Units 3&4
5th Edition
Gary Bass
Natalie Heath
Therese Keane
Anthony Sullivan
Mark Kelly

Senior publisher: Eleanor Gregory
Editor: Scott Vandervalk
Proofreader: Nadine Anderson
Indexer: Bruce Gillespie
Visual designer: James Steer
Cover design: Chris Starr, MakeWork
Text design: Leigh Ashforth, Watershed Art & Design
Permissions researcher: Lyahna Spencer
Production controller: Karen Young
Typeset by: DiacriTech

Any URLs contained in this publication were checked for currency during the production process. Note, however, that the publisher cannot vouch for the ongoing currency of URLs.

Acknowledgements
Extracts from the VCE Applied Computing Study Design (2020–2023), are reproduced by permission, © VCAA. VCE is a registered trademark of the VCAA. The VCAA does not endorse or make any warranties regarding this study resource. Current VCE Study Designs, past VCE exams and related content can be accessed directly at www.vcaa.vic.edu.au

© 2019 Cengage Learning Australia Pty Limited

Copyright Notice
This Work is copyright. No part of this Work may be reproduced, stored in a retrieval system, or transmitted in any form or by any means without prior written permission of the Publisher. Except as permitted under the *Copyright Act 1968,* for example any fair dealing for the purposes of private study, research, criticism or review, subject to certain limitations. These limitations include: Restricting the copying to a maximum of one chapter or 10% of this book, whichever is greater; providing an appropriate notice and warning with the copies of the Work disseminated; taking all reasonable steps to limit access to these copies to people authorised to receive these copies; ensuring you hold the appropriate Licences issued by the Copyright Agency Limited ("CAL"), supply a remuneration notice to CAL and pay any required fees. For details of CAL licences and remuneration notices please contact CAL at Level 11, 66 Goulburn Street, Sydney NSW 2000, Tel: (02) 9394 7600, Fax: (02) 9394 7601
Email: info@copyright.com.au
Website: www.copyright.com.au

For product information and technology assistance,
in Australia call **1300 790 853**;
in New Zealand call **0800 449 725**

For permission to use material from this text or product, please email **aust.permissions@cengage.com**

ISBN 978 0 17 044087 5

Cengage Learning Australia
Level 7, 80 Dorcas Street
South Melbourne, Victoria Australia 3205

Cengage Learning New Zealand
Unit 4B Rosedale Office Park
331 Rosedale Road, Albany, North Shore 0632, NZ

For learning solutions, visit **cengage.com.au**

Printed in Singapore by 1010 Printing International Limited.
1 2 3 4 5 6 7 23 22 21 20 19

Contents

Preface v
About the authors vi
How to use this book vii
Outcomes ix
Problem-solving methodology xiii
Key concepts xvi

Unit 3

Introduction 1

Chapter 1 Data and presentation 2

What is data? 3
Referencing data sources 14
Data types 16
Data structures 19
Data integrity 20
Data validation 24
Data visualisation 25
Designing solutions 37
Design principles 37
Formats and conventions 39
Design tools for data visualisation 46

Chapter 2 Data manipulation and presentation 57

Databases 58
Creating a database structure 59
Spreadsheet tools 84
Data visualisation 89
Testing 96
Review the Outcome's steps 106
Preparing for Unit 3, Outcome 1 112

Chapter 3 Project management and data analysis 113

What is data? 114
Project management 114
Why must we begin with a research question? 124
Primary and secondary data 128
Quantitative and qualitative data 129
Data types and data structures 136
Referencing data sources 137
Data integrity 139
Data security 152
Next steps 158

Chapter 4 Drawing conclusions 165

Continuing Unit 3, Outcome 2 166
Solution specifications 166
Design principles 171
Generating design ideas 176
Design tools 187
Types of infographics and data visualisations 190
Preparing for Unit 3, Outcome 2 204

9780170440875

Unit 4

Introduction 207

Chapter 5 Development and evaluation 208

Data visualisations 209
Procedures and techniques for managing files 214
Functional capabilities of software 220
An effective infographic or data visualisation 221
Manipulating data 234
Formats and conventions 237
Manipulating data with software 238
Verification and validation 243
Testing 244
Evaluating your solution 248
Documenting the progress of projects 250
Assessing your project plan 254
Next steps 255
Preparing for Unit 4, Outcome 1 261

Chapter 6 Information management 262

Networks 263
Threats to data and information 272
Physical and software security controls 275
Managing data on a network 284
Network attached storage and cloud computing 289

Chapter 7 Cyber security measures 297

The importance of data and information to organisations 298
Information management strategies 301
Data security 302
The importance of diminished data integrity in information systems 303
Key legislation relating to data and information 304
Ethics and security practices 314
Resolving legal and ethical tensions 317
Reasons to prepare for disaster 318
Consequences of security failure 328
Evaluating information management strategies 331
Preparing for Unit 4, Outcome 2 338

Index 339

9780170440875

Preface

This fifth edition of *Data Analytics VCE Units 3 & 4* incorporates the changes to the VCAA VCE Applied Computing Study Design that took effect from 2020.

This textbook looks at how individuals and organisations use, and can be affected by, information systems in their daily lives.

We believe that teachers and students require a text that focuses on the **Areas of Study** specified in the **Study Design** and which presents information in a sequence that allows easy transition from theory into practical assessment tasks. We have, therefore, written this textbook so that a class can begin at Chapter 1 and work their way systematically through to the end. Students will encounter material relating to the **key knowledge** dot points for each **Outcome** before they reach the special section that describes the Outcome. The Study Design outlines **key skills** that indicate how the knowledge can be applied to produce a solution to an information problem. These Outcome preparation sections occur regularly throughout the textbook and flag an appropriate point in the student's development for each Outcome to be completed. The authors have covered all key knowledge for the Outcomes from the Data Analytics VCE Units 3 & 4 course.

Our approach has been to focus on the key knowledge required for each school-assessed Outcome, and to ensure that students are well prepared for these; however, there is considerable duplication in the Study Design relating to the knowledge required for many of the Outcomes. We have found that, with an Outcomes approach, we are covering material sometimes several times. For example, knowledge of a problem-solving methodology is listed as key knowledge for many different Outcomes. In these cases, we have tried to provide a general coverage in the first instance, and specifically apply the concept to a situation relevant to the related Outcome on subsequent encounters.

The authors assume that teachers will develop the required key skills with their students within the context of the key knowledge addressed in this textbook and the resources available to them.

We have incorporated a margin column in the text to provide additional information and reinforce key concepts. This margin column also includes activities that relate to the topics covered in the text and considers issues relevant to information systems usage.

Outcome features appear at several points in the book, indicating the nature of the tasks that students undertake in the completion of the school-assessed Outcome. We have listed the steps required to complete the Outcome, together with advice and suggestions for approaching the task. We have also described the output and support material needed for submission. You will also find sample tasks and further advice relating to the Outcomes are available at **https://nelsonnet.com.au**.

The chapters are organised to present the optimum amount of information in the most effective manner. The book is presented in concise, clearly identified sections to guide students through the text. Each chapter is organised into the sections described on pages **vii–viii**.

9780170440875

About the authors

Gary Bass teaches VCE Applied Computing at Year 11 and Year 12 in an online course environment at Virtual School Victoria. Previously, he has taught VCE Physics, as well as developing and delivering middle school ICT courses. Gary has presented at DLTV DigiCON and the annual IT teachers conference on many topics including Pop-up Makerspace; Big Data requires huge analysis – data visualisation; AR + VR = Mixed reality; and Marshall McLuhan-Medium is the message. Gary was selected as an Apple Distinguished Educator (ADE) in 2002 and 2011. In 2016, he was presented with DLTV's IT Leader of the Year award.

Natalie Heath is the eLearning/ICT Leading Teacher at Eltham High School. She has been an IT specialist teacher for nearly 20 years, teaching at all secondary levels including VCE Informatics, IT Applications, Information Processing and Management, and Software Development. She has extensive experience around VCE, having assessed examinations in various subjects for more than two decades. Natalie has also developed many resources for VCE Computing subjects over the years, including trial examinations. She has presented at teacher professional learning conferences as an expert in Unit 3 and 4 subjects and in 2018 was presented with the DLTV's Maggie Iaquinto VCE Computing Educator of the Year award for her services to the VCE Computing teaching community and resource development.

Associate Professor Therese Keane is Deputy Chair of the Department of Education at Swinburne University and has worked in a variety of school settings where she has taught IT in K–12 education as the Director of ICT. Her passion and achievements in ICT in the education and robotics space have been acknowledged by her peers in her receiving numerous national and state awards. Therese has presented seminars and workshops for teachers involved in the teaching of IT. She has written several textbooks in all units of VCE Information Technology. Therese's research interests include the use of technology in education, gender inequalities in STEM-based subjects, robotics in education and computers in schools for teaching and learning purposes. Therese is involved with the FIRST LEGO League as the Championship Tournament Director for Victoria and she is a lead mentor for the RoboCats – a female school student only robotics team that participates in the FIRST Robotic Competition.

Anthony Sullivan is a Curriculum and Learning Specialist at Monash College where he is responsible for creating assessment and learning materials for accounting and computing subjects as part of the Monash University Foundation Year program. Anthony has more than 25 years experience teaching business and computing subjects. He has taught in both government and non-government settings in Australia and taught computing and information technology courses in schools in Asia and the United Kingdom. Anthony has also been a VCE Examination Assessor, a member of the committee that reviewed and wrote the previous Study Design for VCE Computing, and has written a range of commercial resources related to VCE Computing. He has presented at conferences and professional development events and student examination preparation sessions.

9780170440875

How to use this book

KEY KNOWLEDGE

The key knowledge from the VCE Applied Computing Study Design that you will cover in each chapter is listed on the first page of each chapter. The list includes key knowledge specified in the Outcome related to the chapter.

FOR THE STUDENT

Each chapter's opening page includes an overview of that chapter's contents so that you are aware of the material you will encounter.

FOR THE TEACHER

This section is for your teacher and outlines how the chapter fits into the overall study of Data Analytics, and outlines how the material relates to the completion of Outcomes.

CHAPTERS

The major learning material that you will encounter in the chapter is presented as text, photographs and illustrations. The text describes in detail the theory associated with the stated Outcomes of the VCE Applied Computing Study Design in easy-to-understand language. The photographs show hardware, software and other objects that have been described in the text. Illustrations are used to demonstrate concepts that are more easily explained in this manner.

Throughout the chapter, glossary terms are highlighted in bold and you can find their definitions at the end of each chapter, in **Essential terms**.

The **School-assessed Task Tracker** at the bottom of every odd-numbered page provides you with a visual reminder to help you track your progress in the School-assessed Task (SAT), which is derived from Unit 3, Outcome 2 and Unit 4, Outcome 1, so that you can complete all required stages on time.

MARGIN COLUMN

The margin column contains further explanations that support the main text, weblink icons, additional material outside the Study Design and cross-references to material covered elsewhere in the textbook. Issues relevant to Data Analytics that you can discuss with your classmates are also included in the form of 'Think about Data Analytics' boxes.

CHAPTER SUMMARY

The chapter summary at the end of each chapter is divided into two main parts to help you review each chapter.

Essential terms are the glossary terms that have been highlighted throughout the chapter.

The **Important facts** section is a list of summaries, ideas, processes and statements relevant to the chapter, in the order in which they occur in the chapter.

THINK ABOUT DATA ANALYTICS 3.1

Project management tools are useful to find the perfect number of people needed on a task so it is finished as quickly as possible without anyone being idle. Use software to develop a Gantt chart to plan the baking of a cake. Assume you can use as many cooks as you want.

9780170440875

TEST YOUR KNOWLEDGE

Short-answer questions will help you to review the chapter material. The questions are grouped and identified with a section of the text to allow your teacher to direct appropriate questions based on the material covered in class. Teachers will be able to access answers to these questions at **https://nelsonnet.com.au**.

APPLY YOUR KNOWLEDGE

Each chapter concludes with a set of questions requiring you to demonstrate that you can apply the theory from the chapter to more complex questions. The style of questions reflects what you can expect in the end-of-year examination. Teachers will be able to access suggested responses to these questions at **https://nelsonnet.com.au**.

PREPARING FOR THE OUTCOMES

This section appears at points in the course where it is appropriate for you to complete an Outcome task. The information provided describes what you need to do in the Outcome, the suggested steps to be followed in the completion of the task and the material that needs to be submitted for assessment.

NELSONNET

The NelsonNet student website contains:

- multiple-choice quizzes for each chapter, mirroring the VCAA Unit 3 & 4 exam
- additional material such as spreadsheets and infographics.

An open-access weblink page is also provided for all weblinks that appear in the margins throughout the textbook. This is accessible at **https://nelsonnet.com.au**.

The NelsonNet teacher website is accessible only to teachers and it contains:

- answers for the **Test your knowledge** and **Apply your knowledge** questions in the book
- sample School-assessed Coursework (SAC)
- chapter tests
- practice exam.

Please note that complimentary access to NelsonNet and the NelsonNetBook is only available to teachers who use the accompanying student textbook as a core educational resource in their classroom. Contact your sales representative for information about access codes and conditions.

9780170440875

Outcomes

OUTCOME	KEY KNOWLEDGE	LOCATION
Unit 3 Area of Study 1 Outcome 1	Data analytics On completion of this unit the student should be able to respond to teacher-provided solution requirements and designs to extract data from large repositories, manipulate and cleanse data and apply a range of functions to develop software solutions to present findings.	
Data and information	• techniques for efficient and effective data collection, including methods to collect census, Geographic Information System (GIS) data, sensor, social media and weather	pp. 6–10
	• factors influencing the integrity of data, including accuracy, authenticity, correctness, reasonableness, relevance and timeliness	pp. 20–23
	• sources of, and methods and techniques for, acquiring authentic data stored in large repositories	pp. 10–13
	• methods for referencing primary and secondary sources, including American Psychological Association (APA) referencing system	pp. 14–16
	• characteristics of data types	pp. 16–18
Approaches to problem solving	• methods for documenting a problem, need or opportunity	p. 37
	• methods for determining solution requirements, constraints and scope	p. 37
	• naming conventions to support efficient use of databases, spreadsheets and data visualisations	p. 67
	• a methodology for creating a database structure: identifying entities, defining tables and fields to represent entities; defining relationships by identifying primary key fields and foreign key fields; defining data types and field sizes, normalisation to third normal form	pp. 58–66
	• design tools for representing databases, spreadsheets and data visualisations, including data dictionaries, tables, charts, input forms, queries and reports	pp. 46–9, 84
	• design principles that influence the functionality and appearance of databases, spreadsheets and data visualisations	pp. 37–9, 84–93
	• functions and techniques to retrieve required information through querying data sets, including searching, sorting and filtering to identify relationships and patterns	pp. 74–5
	• software functions, techniques and procedures to efficiently and effectively validate, manipulate and cleanse data including files, and applying formats and conventions	pp. 80–2, 88
	• types and purposes of data visualisations	pp. 25–31
	• formats and conventions applied to data visualisations to improve their effectiveness for intended users, including clarity of message	pp. 39–45
	• methods and techniques for testing databases, spreadsheets and data visualisations	pp. 96–106
Interactions and impact	• reasons why organisations acquire data	p. 4
Key skills	• interpret solution requirements and designs to develop data visualisations	pp. 37, 167–71
	• identify, select and extract relevant data from large repositories	pp. 4, 9–13
	• use a standard referencing system to acknowledge intellectual property	p. 14
	• organise, manipulate and cleanse data using database and spreadsheet software	pp. 74–5, 80–8
	• select, justify and apply functions, formats and conventions to create effective data visualisations	pp. 39–45
	• develop and apply suitable validation and testing techniques to software tools used	pp. 24–5

9780170440875

OUTCOME	KEY KNOWLEDGE	LOCATION
Unit 3 Area of Study 2 Outcome 2	**Data analytics: Analysis and design** On completion of this unit the student should be able to propose a research question, formulate a project plan, collect and analyse data, generate alternative design ideas and represent the preferred design for creating infographics or dynamic data visualisations.	
Digital systems	• roles, functions and characteristics of digital system components	p. 148–52
	• physical and software security controls used by organisations for protecting stored and communicated data	p. 152–8
Data and information	• primary and secondary data sources and methods of collecting data, including interviews, observation, querying of data stored in large repositories and surveys	p. 128
	• techniques for searching, browsing and downloading data sets	p. 126
	• suitability of quantitative and qualitative data for manipulation	p. 129–30
	• characteristics of data types and data structures relevant to selected software tools	p. 136–7
	• methods for referencing secondary sources, including the APA referencing system	p. 137–9
	• criteria to check the integrity of data, including accuracy, authenticity, correctness, reasonableness, relevance and timeliness	p. 139–47
	• techniques for coding qualitative data to support manipulation	p. 130–6
Approaches to problem solving	• features of a research question, including a statement identifying the research question as an information problem	p. 124–6
	• functional and non-functional requirements, including data to support the research question, constraints and scope	p. 167–71
	• types and purposes of infographics and dynamic data visualisations	p. 190–8
	• design principles that influence the appearance of infographics and the functionality and appearance of dynamic data visualisations	p. 171–6
	• design tools for representing the appearance and functionality of infographics and dynamic data visualisations, including data manipulation and validation, where appropriate	p. 187–9
	• techniques for generating alternative design ideas	p. 176–85
	• criteria for evaluating alternative design ideas and the efficiency and effectiveness of infographics or dynamic data visualisations	p. 185–7
	• features of project management using Gantt charts, including the identification and sequencing of tasks, time allocation, dependencies, milestones and the critical path	p. 114–23
Interactions and impact	• key legal requirements for the storage and communication of data and information, including human rights requirements, intellectual property and privacy	p. 307–14
Key skills	• frame a research question	p. 124–6
	• analyse and document requirements, constraints and scope of infographics or dynamic data visualisations	pp. 167–171
	• apply techniques for searching, downloading, browsing and referencing data sets	p. 126
	• select and apply design tools to represent the functionality and appearance of infographics or dynamic data visualisations	p. 187–9
	• generate alternative design ideas	p. 176–85

9780170440875

OUTCOME	KEY KNOWLEDGE	LOCATION
	• develop evaluation criteria to select and justify preferred designs	p. 185–7
	• produce detailed designs using appropriate design methods and techniques	p. 187–98
	• propose and apply appropriate methods to secure stored data	p. 152–8
	• create, monitor and modify project plans using software	p. 114–23
Unit 4 Area of Study 1 Outcome 1	**Data analytics: Development and evaluation** On completion of this unit the student should be able to develop and evaluate infographics or dynamic data visualisations that present findings in response to a research question, and assess the effectiveness of the project plan in monitoring progress.	
Digital systems	• procedures and techniques for handling and managing files, including archiving, backing up, disposing of files and security	pp. 214–20
	• the functional capabilities of software to create infographics and dynamic data visualisations	p. 220
Approaches to problem solving	• characteristics of information for educating targeted audiences, including age appropriateness, commonality of language, culture inclusiveness and gender	p. 221–9
	• characteristics of efficient and effective infographics and dynamic data visualisations	p. 221–34
	• functions, techniques and procedures for efficiently and effectively manipulating data using software tools	pp. 234, 238
	• techniques for creating infographics and dynamic data visualisations	pp. 238–42
	• techniques for validating and verifying data	p. 243
	• techniques for testing that solutions perform as intended	pp. 244–8
	• techniques for recording the progress of projects, including adjustments to tasks and timeframes, annotations and logs	pp. 250–3
	• strategies for evaluating the effectiveness of infographics and dynamic data visualisations solutions and assessing project plans	pp. 248–50
Key skills	• monitor, modify and annotate the project plan as necessary	p. 251
	• propose and implement procedures for managing files	pp. 214–20
	• select and apply software functions, conventions, formats, methods and techniques to develop infographics or dynamic data visualisations	pp. 237–8
	• select and apply data validation and testing techniques, making any necessary modifications	pp. 243–8
	• apply evaluation criteria to evaluate the efficiency and effectiveness of infographics or dynamic data visualisations solutions	pp. 248–50
	• assess the effectiveness of the project plan in managing the project	pp. 254–5

9780170440875

OUTCOME	KEY KNOWLEDGE	LOCATION
Unit 4 Area of Study 2 Outcome 2	**Cybersecurity: Data and information security** On completion of this unit the student should be able to respond to a teacher-provided case study to investigate the current data and information security strategies of an organisation, examine the threats to the security of data and information, and recommend strategies to improve current practices.	
Digital systems	• characteristics of wired, wireless and mobile networks	pp. 263–71
	• types and causes of accidental, deliberate and events-based threats to the integrity and security of data and information used by organisations	pp. 272–5
	• physical and software security controls for preventing unauthorised access to data and information and for minimising the loss of data accessed by authorised and unauthorised users	pp. 275–84
	• the role of hardware, software and technical protocols in managing, controlling and securing data in information systems	pp. 284–9
	• the advantages and disadvantages of using network attached storage and cloud computing for storing, communicating and disposing of data and information	pp. 289–90
Data and information	• characteristics of data that has integrity, including accuracy, authenticity, correctness, reasonableness, relevance and timeliness	pp. 20–3
Interactions and impacts	• the importance of data and information to organisations	p. 298
	• the importance of data and information security strategies to organisations	p. 302
	• the impact of diminished data integrity in information systems	pp. 303–4
	• key legislation that affects how organisations control the collection, storage, communication and disposal of their data and information: the *Health Records Act 2001*, the *Privacy Act 1988* and the *Privacy and Data Protection Act 2014*	pp. 304–13
	• ethical issues arising from data and information security practices	pp. 314–17
	• strategies for resolving legal and ethical issues between stakeholders arising from information security practices	pp. 317–18
	• reasons to prepare for disaster and the scope of disaster recovery plans, including backing up, evacuation, restoration and test plans	pp. 318–27
	• possible consequences for organisations that fail or violate security measures	p. 328
	• criteria for evaluating the effectiveness of data and information security strategies	pp. 331–3
Key skills	• analyse and discuss the current data and information security strategies used by an organisation	pp. 298–302
	• propose and apply criteria to evaluate the effectiveness of current data and information security strategies	p. 302
	• identify and evaluate threats to the security of data and information	pp. 272–5, 302
	• identify and discuss possible legal and ethical consequences of ineffective data and information security strategies	p. 328
	• recommend and justify strategies to improve current data and information security practices	p. 328

Reproduced from the VCE Applied Computing Study Design (2020–2023) © VCAA; used with permission.

9780170440875

Problem-solving methodology

When an information problem exists, a structured problem-solving methodology is followed to ensure that the most appropriate solution is found and implemented. For the purpose of this course, the problem-solving methodology has four key stages: analysis, design, development and evaluation. Each of these stages can be further broken down into a common set of activities. Each unit may require you to examine a different set of problem-solving stages. It is critical for you to understand the problem-solving methodology because it underpins the entire VCE Applied Computing course.

Reproduced from the VCE Applied Computing Study Design (2020 – 2023) © VCAA; used with permission.

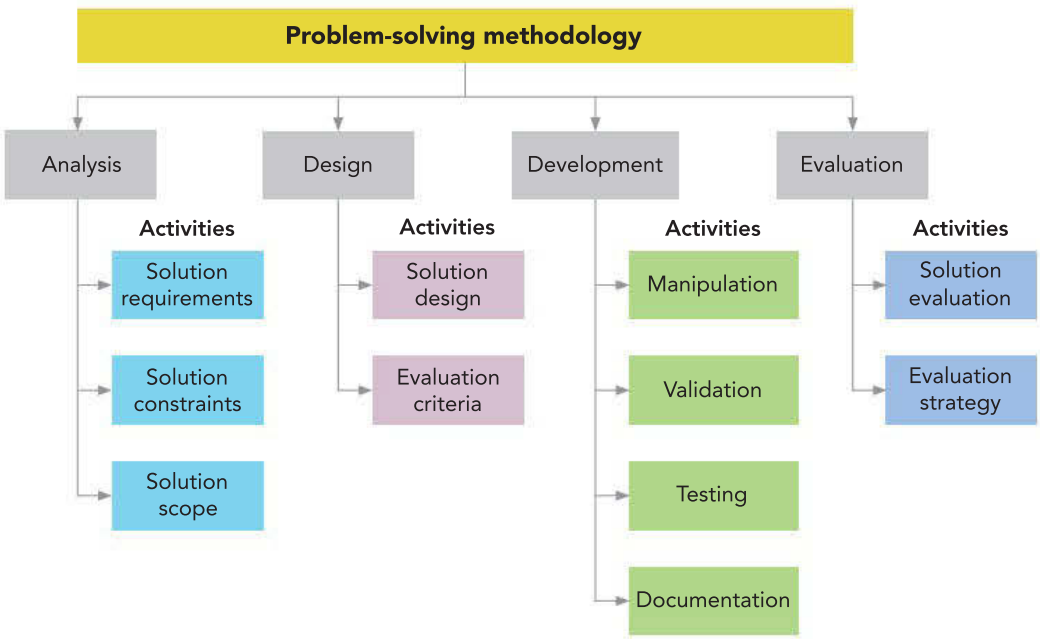


FIGURE 1 The four stages of the problem-solving methodology and their key activities

Analyse the problem

The purpose of analysis is to establish the root cause of the problem, the specific information needs of the organisation involved, limitations on the problem and exactly what a possible solution would be expected to do (the scope). The three key activities are:

1. identifying solution requirements – attributes and functionality that the solution needs to include, information it must produce and data needed to produce this information
2. establishing solution constraints – the limitations on solution development that need to be considered; constraints are classified as economic, technical, social, legal and related to usability
3. defining the scope of the solution – what the solution will and will not be able to do.

9780170440875

Design the solution

During the design stage, several alternative design ideas based on both appearance and function are planned and the most appropriate of these is chosen. Criteria are also created to select the most appropriate ideas and to evaluate the solution's success once it has been implemented. The two key design activities include the following.

1. Creating the solution design – it must clearly show a developer what the solution should look like, the specific data required, and how its data elements should be structured, validated and manipulated. Tools typically used to represent data elements could include data dictionaries, data structure diagrams, input–process–output (IPO) charts, flowcharts, pseudocode and object descriptions. The following tools are also used to show the relationship between various components of the solution: storyboards, site maps, data flow diagrams, structure charts, hierarchy charts and context diagrams. Furthermore, the appearance of the solution, including elements like a user interface, reports, graphic representations or data visualisations, needs to be planned so that overall layout, fonts and their colours can be represented. Layout diagrams and annotated diagrams (or mock-ups) usually fulfil this requirement. A combination of tools from each of these categories will be selected to represent the overall solution design. Regardless of the visual or functional aspects of a solution design at this stage, a design for the tests to ultimately ensure the solution is functioning correctly must also be created.
2. Specifying evaluation criteria – during the evaluation stage, the solution is assessed to establish how well it has met its intended objectives. The criteria for evaluation must be created during the design stage so that all personnel involved in the task are aware of the level of performance that will ultimately determine the success or otherwise of the solution. The criteria are based on the solution requirements identified at the analysis stage and are measured in terms of efficiency and effectiveness.

Develop the solution

The solution is created by the developers during this stage from the designs supplied to them. The 'coding' takes place, but also checking of input data (validation), testing that the solution works and the creation of user documentation. The four activities involved with development include the following.

1. Manipulating or coding the solution – the designs are used to build the electronic solution. The coding will occur here and internal documentation will be included where necessary.
2. Checking the accuracy of input data by way of validation – manual and electronic methods are used; for example, proofreading is a manual validation technique. Electronic validation involves using the solution itself to ensure that data is reasonable by checking for existence, data type and that it fits within the required range. Electronic validation, along with any other formulas, always needs to be tested to ensure that the solution works properly.

9780170440875

3 Ensuring that a solution works through testing – each formula and function, not to mention validation and even the layout of elements on the screen, needs to be tested. Standard testing procedures involve stating what tests will be conducted, identifying test data, stating the expected result, running the tests, reporting the actual result and correcting any errors.

4 Documentation allowing users to interact with (or use) the solution – while it can be printed, in many cases it is now designed to be viewed on screen. User documentation normally outlines procedures for operating the solution, as well as generating output (like reports) and basic troubleshooting.

Evaluate the solution

Sometimes after a solution has been in use by the end-user or client, it needs to be assessed or evaluated to ensure that it has been successful and does actually meet the user's requirements. The two activities involved in evaluating a solution include the following.

1 Evaluating the solution – providing feedback to the user about how well the solution meets their requirements or needs or opportunities in terms of efficiency and effectiveness. This is based on the findings of the data gathered at the beginning of the evaluation stage when compared with the evaluation criteria created during the design stage.

2 Working out an evaluation strategy – creating a timeline for when various elements of the evaluation will occur and how and what data will be collected (because it must relate to the criteria created at the design stage).

9780170440875

Key concepts

Within each VCE Applied Computing subject are four key concepts the purpose of which is to organise course content into themes. These themes are intended to make it easier to teach and make connections between related concepts and to think about information problems. Key knowledge for each Area of Study is categorised into these key concepts, but not all concepts are covered by each Area of Study. The four key concepts are:

1. digital systems
2. data and information
3. approaches to problem solving
4. interactions and impact.

Digital systems focus on how hardware and software operate in a technical sense. This also includes networks, applications, the internet and communication protocols. Information systems have digital systems as one of their parts. The other components of an information system are people, data and processes.

Data and information focuses on the acquisition, structure, representation and interpretation of data and information in order to elicit meaning or make deductions. This process needs to be completed in order to create solutions.

Approaches to problem solving focuses on thinking about problems, needs or opportunities and ways of creating solutions. Computational, design, and systems thinking are the three key problem-solving approaches.

Interactions and impact focuses on relationships that exist between different information systems and how these relationships affect the achievement of organisational goals and objectives. Three types of relationships are considered:

1. people interacting with other people when collaborating or communicating with digital systems
2. how people interact with digital systems
3. how information systems interact with other information systems.

This theme also looks at the impact of these relationships on data and information needs, privacy, and personal safety.

9780170440875

Unit 3

INTRODUCTION

In Unit 3 of Data Analytics, you will use data that you have effectively and efficiently identified and extracted. You will consider the integrity and source of the data and make sure that you have correctly referenced the data using the American Psychological Association (APA) referencing system. You will manipulate this data to create data visualisations and infographics. To do this, you will use databases, spreadsheets and data manipulation software.

You will propose a research question and then collect data to answer this research question. You will use a range of methods to analyse this data. You will use all the stages of the problem-solving methodology (PSM) to prepare a project plan. This will complete the first half of the School-assessed Task (SAT) (Unit 3, Outcome 2). The second half of the SAT will be completed in Unit 4 (Unit 4, Outcome 1).

Area of Study 1 – Data analytics

OUTCOME 1 In this Outcome, you will respond to teacher-provided solution requirements and designs. You will extract data from large data repositories. You will manipulate and cleanse the data and use database, spreadsheet and data manipulation software to present your findings in the form of a data visualisation.

Area of Study 2 – Data analytics: Analysis and design

OUTCOME 2 In this Outcome, you will propose a research question and then collect data to answer this research question. You will use a range of methods to analyse this data. You will use all the stages of the problem-solving methodology to prepare a project plan. This will complete the first half of the School-assessed Task (SAT) (Unit 3, Outcome 2). The second half of the SAT will be completed in Unit 4 (Unit 4, Outcome 1).



iStock.com / gremlin

CHAPTER

1 Data and presentation

KEY KNOWLEDGE

After completing this chapter, you will be able to demonstrate knowledge of:

Data and information

- techniques for efficient and effective data collection, including methods to collect census, Geographic Information System (GIS) data, sensor, social media and weather
- factors influencing the integrity of data, including accuracy, authenticity, correctness, reasonableness, relevance and timeliness
- sources of, and methods and techniques for, acquiring authentic data stored in large repositories
- methods for referencing primary and secondary sources, including American Psychological Association (APA) referencing system
- characteristics of data types

Approaches to problem solving

- methods for documenting a problem, need or opportunity
- methods for determining solution requirements, constraints and scope
- design tools for representing databases, spreadsheets and data visualisations, including data dictionaries, tables, charts, input forms, queries and reports
- design principles that influence the functionality and appearance of databases, spreadsheets and data visualisations
- formats and conventions applied to data visualisations to improve their effectiveness for intended users, including clarity of message

Interactions and impact

- reasons why organisations acquire data.

Reproduced from the VCE Applied Computing Study Design (2020–2023) © VCAA; used with permission.

FOR THE STUDENT

If you can imagine the sheer amount of data that is generated every day, you might also be able to imagine that there is someone, somewhere, who is looking through a mountain of data searching for meaning. Data visualisation is the process by which we take large amounts of data and process it into effective graphical representations that will meet the needs of users or clients. These representations can take the form of charts, graphs, spatial relationships and network diagrams.

In some cases, the data visualisation might involve interactivity and the inclusion of dynamic data that allows the user to deduce further meaning from the visualisation. This chapter will cover the definitions of data and information, the various ways in which data can be acquired and referenced and how to check that data is reliable enough to be used to generate useful information. This chapter will then look at the many types of data visualisations and the design tools that could be used to help plan their use.

FOR THE TEACHER

This chapter introduces students to the knowledge and skills needed to use software tools to access authentic data from repositories and present the information in a visual form. It covers data types, data integrity and citing references before covering a range of data visualisation tools and their purposes.

The key knowledge and skills are based on Unit 3, Area of Study 1. If a data visualisation is effective, it reduces the effort needed by readers to interpret information. This chapter takes students through the different types of visualisations. This chapter, combined with Chapter 2, will form a foundation for the Unit 3, Outcome 1 School-assessed Coursework (SAC). Much of this will be applied when students work on their School-assessed Task (SAT).



iStock.com/monsitj

9780170440875

What is data?

Data is made up of facts and statistics. Raw facts have no context to them, so you cannot make much sense of them, or give them any meaning. To understand and make meaning of data, you need to process or manipulate it, converting it into something useful: **information**.

Data consists of raw, unorganised facts, figures and symbols fed into a computer during the input process. Data can also refer to ideas or concepts before they have been refined. In addition to text and numbers, data also includes sounds and images (that are both still and moving). Organisations are collecting this data in vast quantities every day. It allows them to plan day-to-day operations, make better business decisions as well as to better understand their customers.

There are several ways in which this data is gathered by organisations. Popular methods include analysing comments on social media, tracking activity on product websites, placing cookies on customer computers and tracking IP addresses.

Organisations can then use data for specific purposes such as targeting customers with advertising tailored to their interests, developing new products and improving existing ones, and even protecting data (for instance, when banks analyse your credit card usage patterns to identify potential fraudulent transactions). There is no doubt that understanding the customer remains a key need of any organisation.

The potential value of data cannot be fully unlocked without processing it into information. Marked ballot papers after an election hold a great deal of data, and thus considerable potential value, but in their raw form, they hold little value. They need to be processed through counting and grouping the ballot papers into their electorates before they become useful information – that is, election results.

Datum is the singular form of data, which is technically plural. Today, nearly everyone uses *data*, the plural form, for both singular and plural.

A cookie is a small file that a web server stores on a user's computer. Cookies typically contain data about the user, such as their email address and browsing preferences. The cookie is sent to the computer when a website is browsed and stored on the computer's hard disk. The next time the website is visited, the browser retrieves the cookie from the hard disk and sends the data in the cookie to the website. Cookies are not viruses because they cannot be executed or run, and they cannot replicate themselves; however, they can be misused as spyware. Cookies can be used to track people, which leads to privacy issues.

a b c



Getty Images/Paul Kane



Getty Images/AFP/WILLIAM WEST

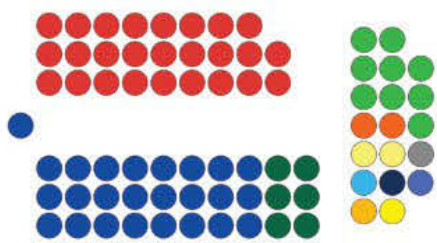


FIGURE 1.1 a Raw data in the form of ballot papers are b grouped and counted to produce c election results.

In this chapter, you will learn more about data, data collection and how information is formed and visualised through manipulation. As part of Unit 3, Outcome 1, you will need to collect data and manipulate it using both a relational database and spreadsheet software to create meaningful data visualisations. These manipulations will be covered in detail in Chapter 2.

We will also discuss the data you need to collect, in terms of how it should be treated and manipulated, and how limiting factors involving **constraints** and scope can affect the data.

Data should be gathered from reputable sources, so we will cover how you can acquire data through existing sources, and then measure suitability and integrity by questioning how the data was acquired, such as through surveys, interviews or observation. We will also discuss how to reference those sources properly.

The IP addressing standard – four numbers between 0 and 255 separated by full stops – defines a mechanism to provide a unique address for each computer on a network.

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

Once you have gathered all of this data, you need to store it, protect it and understand what type of data it is. We will discuss data integrity and how to maintain it through measures such as timeliness, accuracy, authenticity and relevance. This is important because you need to maintain the integrity of the data collected for your Outcome so that your final product can be considered reliable.

Why organisations acquire data

Organisations depend on data in order to function. They use it to keep track of stock levels, client details, employee details, rosters, finances and records of their work. They also use data analytics to make predictions about all aspects of their business.

If an organisation were to lose all of its data, it would suffer greatly. The organisation would not be able to keep track of any finances, it would lose track of its client-base and, if the general public became aware of the data loss, it would suffer a loss of reputation. This would potentially result in fewer clients and the possibility of legal action and closure of the business.

Information needs

When clients or users require particular information, and no system currently exists that provides the information, then an **information need** has been identified.

This could be due to an existing information problem (an organisation worried about declining sales), an identified need (park rangers needing a method to communicate weather conditions on total fire ban days), or an opportunity (currently no list of driving instructors in Victoria exists).

When an information need has been identified, one process used to create a solution that will meet the needs of the clients or users is the **problem-solving methodology (PSM)**.

Data acquisition

Acquisition is when raw data is gathered from the world outside the **information system**. First-hand, or **primary data**, may be acquired manually, via surveys, interviews or observation or it may be acquired electronically. Electronic acquisition can be completed in many ways: through cameras, people inputting data manually, sensors detecting something such as movement, or it may be acquired through other electronic means (for example, if using a keylogger or scanner).

Data can also be acquired by locating repositories of data that already exist, often online, that has been compiled by someone else or another organisation. If it has not been collected by the organisation directly, or if it has been manipulated or summarised in any way, then it is considered **secondary data**. This form of data can save a lot of time, but you must ensure that the data has come from a trustworthy source. It is important to know how this data was collected, by whom and if there are any reasons to doubt its reliability.

Trusting that data has been collected and stored in a secure manner is discussed on page 20 under the 'Data integrity' heading.

Primary and secondary data

Data that has not been filtered by interpretation or evaluation is called primary data. Often, these are facts that you, the researcher, have collected directly to answer a specific question, but it may also be old data that has never been given proper scrutiny before. The lack of previous interpretation is what categorises something as primary data.

9780170440875

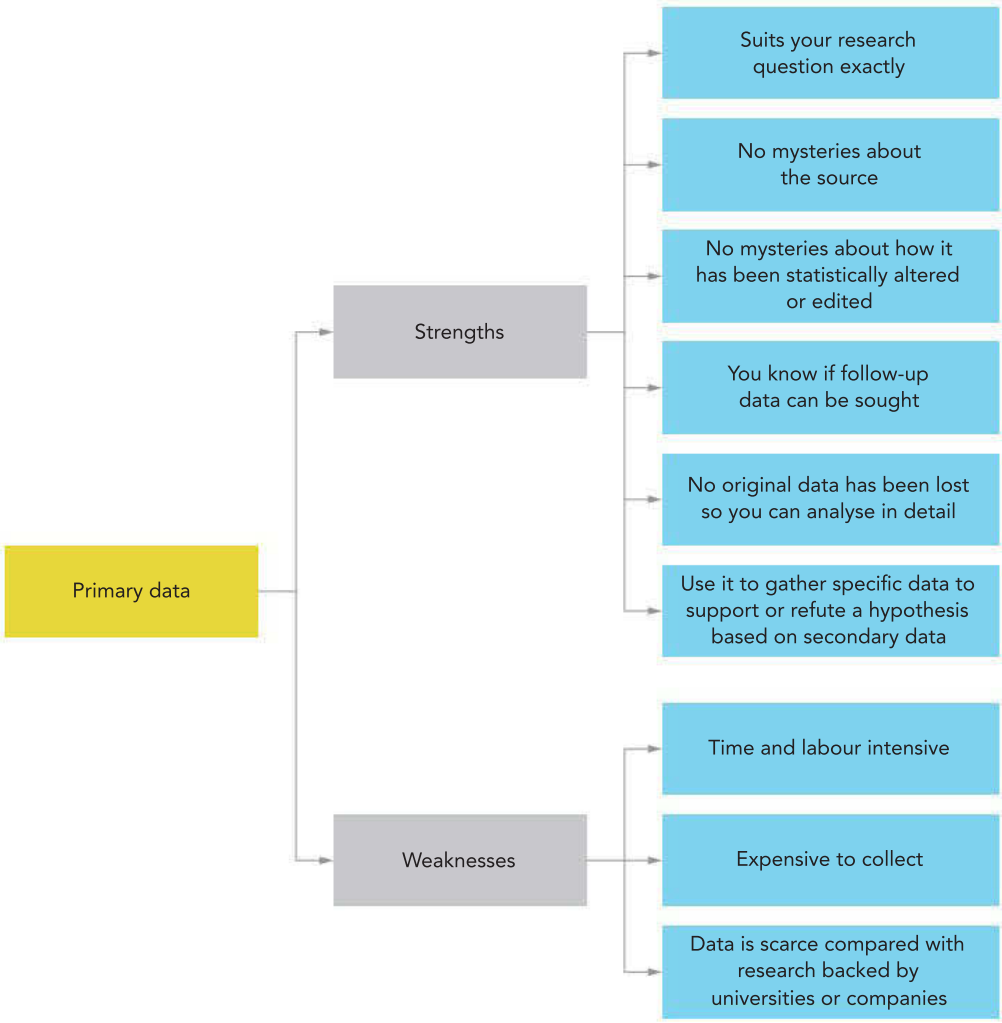


FIGURE 1.2 Strengths and weaknesses of primary data

Journal articles, results of questionnaires, diaries and letters, emails, internet postings, speeches, audio and video recordings (if unedited), official documents and photos may all be primary data. Therefore, when *you*, as the researcher, need to interpret or evaluate unedited data that you are using in your Outcome, it is primary data.

Researchers also collect (new) primary data using interviews, focus groups, surveys, experiments, observations and measurements.

Secondary data differs from primary data because it has been collected and interpreted by someone *other* than the researcher. The collectors of the data could be other researchers, or government departments, or any of a variety of sources: encyclopaedias and other books, biographers, conductors of polls and surveys, journalists, newspapers and magazines, the Australian Bureau of Statistics, internet posts, databases and so on. When using secondary data, it is especially important to consider whether both the data and its interpretation are from a reputable source. Therefore, when you collect data for your Outcome that has been interpreted by someone else already, you should consider this to be secondary data.

9780170440875

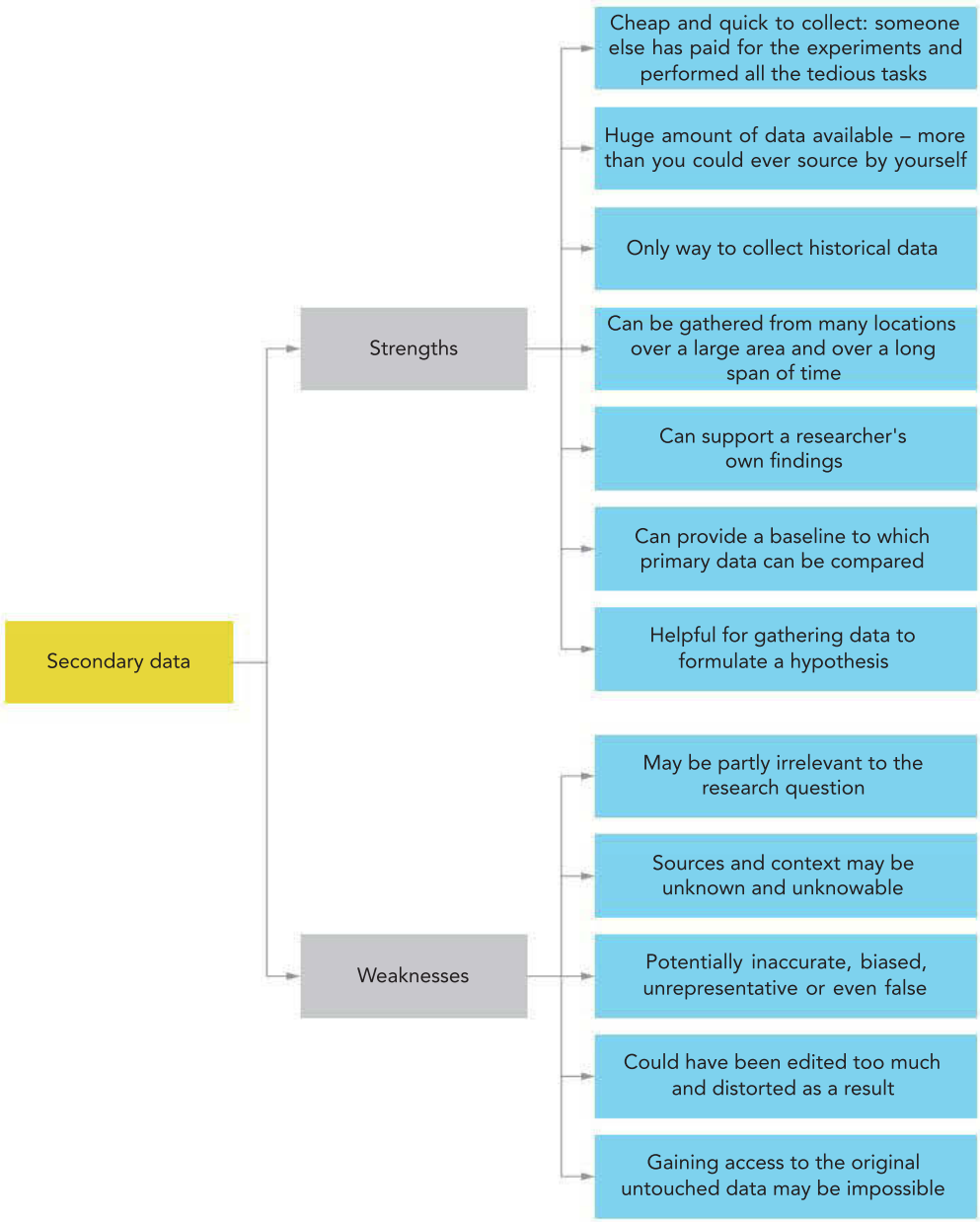


FIGURE 1.3 Strengths and weaknesses of secondary data

Techniques to collect data

Techniques to collect data will also be covered later when you begin working on your SAT for Unit 3, Outcome 2.

Ideally, data should be collected both **efficiently** (not wasting time, cost or effort) and **effectively** (in a way that ensures data is complete, usable, accurate and current). How the data is collected can affect the quality or integrity of the data. Data can be collected by using surveys, interviewing participants or observation. Each method has its own benefits and drawbacks; you need to know how the data you are using has been collected in order to understand the impact it will have upon the quality of your first SAC.

9780170440875

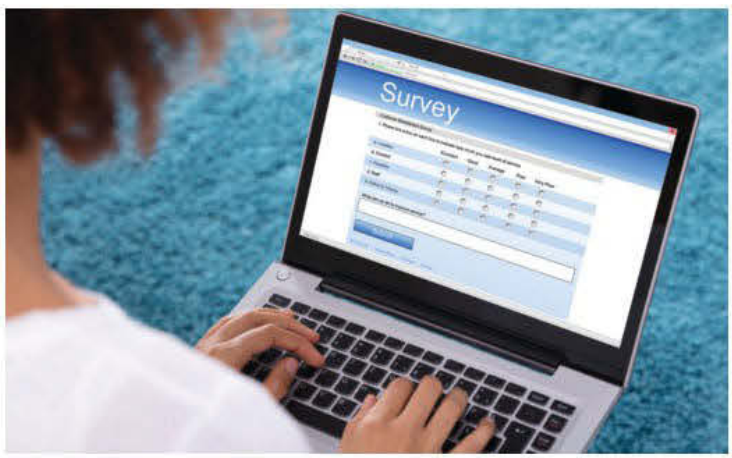
Survey

Surveys are a fast and relatively cheap way to gather large amounts of data and feedback. They can be administered in many different ways – online (having users enter their own data makes things much easier and quicker), on paper (circling numbers, ticking boxes, writing short responses) or verbally in person or over the telephone. The questions in a survey remain identical for each person completing it, so that if any further clarification is required, this cannot be done, especially if the survey is anonymous. If it becomes apparent that extra questions need to be added, it is probably too late to do so – once produced, a survey is fixed.

For example, if a respondent misunderstood a question, or if the response given contradicts an earlier response, then the quality of data might be affected. There is no way of gauging whether or not people are being honest in their responses as with a face-to-face interview, which might allow for tone and body language to make points of view clearer.

Question types are limited to:

- checkbox for Boolean data: yes/no or true/false questions only
- scaled responses: Likert scales ask respondents to select how they feel about a particular statement, asking for a number from 1–5 (or any scale)
- closed questions: asking for a response from a set number of options
- open questions: respondents give worded responses with no limitations; this can give more detail, but it is more difficult to use this data when collecting from a large number of people.



Shutterstock.com/Andrey_Popov

FIGURE 1.4 Conducting surveys online saves time and effort because they avoid added manual data entry.

Interview

Interviews take place with two or more people in real time. They can be conducted face-to-face, via video or telephone, and with individuals or small groups. Like surveys, the questions are usually written in advance so that responses can be compared to get the big picture from the data collected. Unlike surveys, respondents can request clarification if they do not understand a question, or the interviewer can probe for more detail if they think that it is appropriate. For example, an interviewer might ask the respondent to provide an example or ask why they said something. Interviews are more costly, in that they cannot be deployed to as many people or as quickly as surveys, but the quality of data is usually better.

9780170440875



Shutterstock.com/GaudiLab

FIGURE 1.5 Interviewing allows for follow-up and clarification of questions.

Observation

Observations are the most costly data-collection technique when collecting data about real-time events or existing processes. They involve watching and taking notes in real time as an impartial observer. This is far slower than the previous two techniques, but results in far richer and more genuine data. This kind of data collection is better suited to studies that do not require large amounts of data. Observations are generally a good idea to combine with surveys and interviews to allow more information to be collected. For example, if a respondent's body language in an interview indicated that they were unsure of their answer, the interviewer could request more detail around the response to gather more usable data.

Sensor

Sensors or data loggers are devices used to detect characteristics of the environment around them. This may include temperature, humidity, light levels, motion, touch, and the amount of gases in the air. Sensors are always connected to other electronics that can interpret the electrical signals they generate. Sensors can be connected together to provide an overall snapshot of a particular environment (as with a weather station) and even set-up to transmit



iStock.com/Phuchit

FIGURE 1.6 Sensors can be used to measure air quality.

data from remote locations (including other planets). It is important that data collected from sensors is validated to reduce anomalies and stored securely to reduce the possibility of data loss.

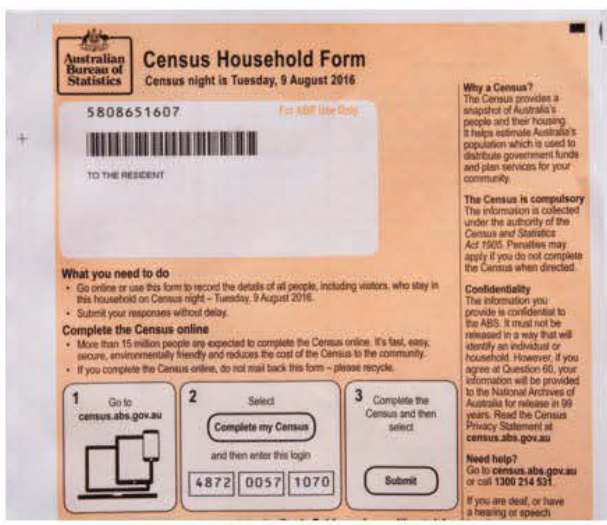
Sensors collect data electronically. Once installed, they can monitor and transfer data to storage without human intervention. They can operate 24 hours a day, 7 days a week, and they can gather data on weather, movement, traffic speeds on roads, levels of light, noise or pollution, or numbers of people or cars entering or leaving facilities, or almost anything else.

Methods applied to specific data collections

The following data repositories have collected vast amounts of data in various ways. The organisations and their data will be discussed later in this chapter. Considering the sheer volume of data being collected by organisations, it would be impossible for a human to record it either efficiently or effectively. Therefore, there are several automated processes commonly used to gather data. Common types of data to be gathered include census, **Geographic Information System (GIS)** data, sensor, social media and weather data.

Census

Every five years, the Australian Bureau of Statistics (ABS) conducts the Census of Population and Housing. This is one of the largest data collecting activities in Australia. It is designed to provide a demographic snapshot of Australian society. In previous years, households were asked to complete a paper survey booklet containing questions about various characteristics of the people living in a house on a particular night (Census Night). The 2016 Census was the first time census data could be entered via an online portal. Households were provided with login details for authentication. Responses to questions were validated electronically and deidentified. While the collection of census data was easier in 2016 when compared with the paper-based forms, the data entry process was more vulnerable to interference, as was demonstrated when hackers caused the portal to be shut down temporarily. Those who were unable to complete the census online were given extra days to complete the task. It is a legal requirement for each household to complete the Census, and the honest, accurate completion of nearly every question is required.


Australian Bureau of Statistics

Census Household Form

Census night is Tuesday, 9 August 2016

5808651607

TO THE RESIDENT

What you need to do

- Go online or use this form to record the details of all people, including visitors, who stay in this household on Census night – Tuesday, 9 August 2016.
- Submit your responses without delay.

Complete the Census online

- More than 15 million people are expected to complete the Census online. It's fast, easy, secure, environmentally friendly and reduces the cost of the Census to the community.
- If you complete the Census online, do not mail back this form – please recycle.

1 Go to census.abs.gov.au

2 Select Complete my Census and then enter this login 4872 0057 1070

3 Complete the Census and then select Submit

Why a Census?
The Census provides a snapshot of Australia's people and their housing. It helps estimate Australia's population which is used to distribute government funds and plan services for your community.

The Census is compulsory
The information is collected under the authority of the *Census and Statistics Act 1905*. Penalties may apply if you do not complete the Census when directed.

Confidentiality
The information you provide is confidential to the ABS. It must not be released in a way that will identify an individual or household. However, if you agree at Question 60, your information will be provided to the National Archives of Australia for release in 99 years. Read the Census Privacy Statement at census.abs.gov.au

Need help?
Go to census.abs.gov.au or call 1300 214 531.

If you are deaf, or have a hearing or speech

Alamy Stock Photo / chris24

FIGURE 1.7 The ABS census provides a snapshot of Australians and Australian households every five years.

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

Weather

The Bureau of Meteorology (BOM) uses sensors to collect data on all aspects of weather and climate. Data is automatically gathered and stored and is made available on the BOM website. The sensors are not subject to personal bias or human error, and unless they malfunction their accuracy and reliability is very high.

The National Oceanic and Atmospheric Administration (NOAA) and the National Aeronautics and Space Administration (NASA) in the United States also use sensors to gather climate data, which is available from their websites.

Social media

Many social media sites generate data from their users. This data, sometimes only available through subscription, summarises how far the account has reached, and how many views and interactions have happened. It is not always easy to locate the raw data.

Social media sites also take part in data mining – storing data about its users. This data is gathered in order to learn about users for, among other things, targeted advertising.

Interaction with popular social media platforms such as Facebook, Twitter and LinkedIn provides a range of measurable data elements. This data comes from interactions such as clicks, comments, likes, shares and conversations. Social media users can also be broken down further by location and language preferences. This data can reveal the success of a marketing campaign or evaluate customer impressions of a product.

Geographic Information System (GIS)

Geographic Information Systems use sensors to record data about Earth. They capture the data, store it, carry out manipulation to create useful information and then present it in easily understood ways. A GIS stores data on geographical features as well as their characteristics. Features are usually classified as points, lines or areas. Data might also be stored as images. City data on a map might be stored as points, road data as lines, and boundaries as areas, while aerial photographs or scanned maps could be stored as raster images. The ABS presents a lot of their data as GIS data.

Data sources

Finding a relevant **data source** takes thought, judgement and care. A great deal of data is available online. The trick is to find the data that is worth using and to use it correctly. While Chapter 3 deals with the forming of a research question, this chapter is about the nature of data itself.

It is important that you investigate both primary and secondary data sources. This will be discussed in much more detail as you prepare for the next Outcome.

The source of data is very important. Treat data without an identified source with care and take necessary steps to resolve matters of **authenticity**. When the source of the data is known, it is more reliable. Without this, your trust would be blind because you would not be able to contact the data's creator – the data could even be a work of fiction. Putting complete trust in anonymous data is especially unwise. For example, if you find survey results on social media, it is important to check to see if it is supported by substantiated facts and that the data has not been made up by somebody.

However, if you know how data has been collected, statistically manipulated, edited and/or abridged, you can interpret the information more wisely. A reputable data source

with authority is more likely to provide high-quality data. Be wary of data given by people who *are* experts, but *not* necessarily experts in the relevant field. The opinion of a champion footballer on politics, or an actor on climate change, are no better than any other person's opinion.

There are hundreds of data sites in Victoria and Australia on many topics. Choose a question that interests you. Search online for key words relevant to your question and add 'data set'.

You will probably find data that will inspire many reasonable and interesting research questions.

Data from government organisations

Data that is collected by organisations is sometimes made available to the public, often online. Many government bodies, including Australian federal and state governments, make some data sets available. For Unit 3, Outcome 1, you will be using some of this data. It is important to know about these organisations since you need to be able to trust them as data sources.

Australian Bureau of Statistics (ABS)

The Australian Bureau of Statistics (ABS) is the statistical agency of the federal government. The ABS provides statistics on a wide range of economic, environmental and social issues, for use by governments and the community. Data sets are available on a diverse range of topics, from foreign trade to agriculture, sporting facilities and crime statistics. In fact, the ABS website provides an endless supply of open data, downloadable and ready for manipulation.

Australian Bureau of Statistics

DataVic

DataVic offers public data sets from many locations, groups and topics, including a national public toilet map, overseas arrivals and departures, taxation statistics and even Victorian ice-skating centres.

DataVic

Commonwealth Scientific and Industrial Research Organisation (CSIRO)

The CSIRO is an Australian scientific organisation that collects and stores scientific data. It manages national scientific research and focuses on challenges that face both Australia and the world. The CSIRO collaborates with many leading organisations worldwide to solve challenges using innovation around science and technology, and is committed to shaping the future to improve communities, economies and the planet.

CSIRO

Bureau of Meteorology (BOM)

The Bureau of Meteorology is Australia's national weather, climate and water agency. The BOM collects a wide range of climatic data in order to make regular forecasts, issue warnings and offer advice with regards to the weather. Much of the data that it collects is available for public use.

Bureau of Meteorology

Weather data is collected by a range of devices including thermometers (temperature), radar systems (used to create maps of rain and snow and measure rain cloud movement), barometers (air pressure), rain gauges (how much rain falls), wind vanes (wind speed and direction), transmissometers (visual range), and hygrometers (humidity and water vapour). These devices work in real time to provide current weather information as well as assist with weather prediction.

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

The Bureau of Meteorology operates a large number of weather stations throughout Australia that have a combination of these devices. Australian weather data is also shared with other countries, and vice versa, in order to establish global weather patterns. Analysis of weather data is normally done on a supercomputer.



Reproduced by permission of Bureau of Meteorology, © 2019 Commonwealth of Australia

FIGURE 1.8 The BOM's climate data online

NASA

National Aeronautics and Space Administration (NASA)

The United States' National Aeronautics and Space Administration (NASA) website has large repositories of data available to download. Its data and research are focused not only on space, but geographic and climate data for Earth as well. For example, much of the data presented about global warming and pattern trends worldwide comes from NASA.

1.1 THINK ABOUT DATA ANALYTICS

Create a list of four other organisations in Australia that provide data sets for public use. Each organisation can be a government department or private organisation.

Not-for-profit organisations

Organisation for Economic Co-operation and Development (OECD)

The Organisation for Economic Co-operation and Development (OECD) is a group of countries working together. The aim is to use its wealth of information on a broad range of topics to help governments foster prosperity and fight poverty through economic growth and financial stability. The OECD helps to ensure the environmental implications of economic and social development are taken into account. The OECD collects global data, by country, across many areas deemed important to making lives better for people, including gross domestic product (GDP), education levels, mortality rates, health data, financial data and much more. Investigate Gapminder online (see weblink) to engage with some of their data.

Gapminder
Gapminder is a not-for-profit organisation that specialises in data visualisation of large data sets from reputable sources including OECD, WHO (World Health Organization), UNAIDS, International Labour Organization, UNESCO and the United Nations. You can choose your own data sets and create stunning visualisations.

Other collections of data

Social media

Social media sites have in-built tools to help users locate data to track their impact. 'Likes' and 'shares' are attached to each post, as are comments. Figure 1.9 shows some Twitter feedback on an individual tweet.

Sensor data

Sensor data is anything that is collected via sensors. These can include road (traffic) cameras, video surveillance, toll data, Bluetooth capture boxes or something that you set-up yourself that you could use to potentially aid your research in your Unit 3, Outcome 2 SAT.

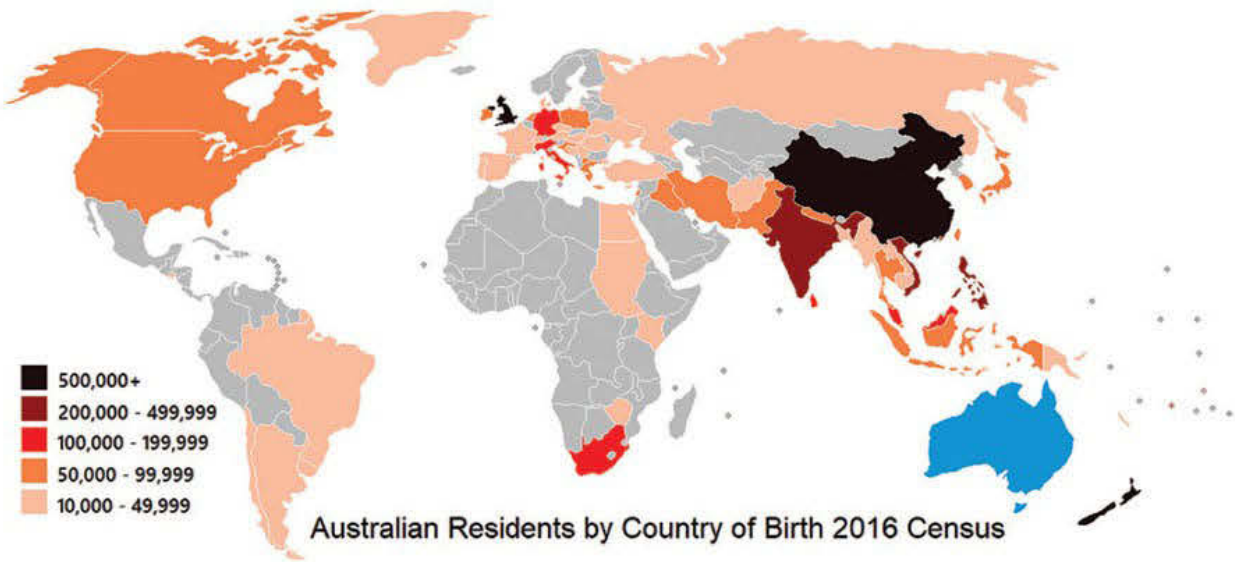
This type of data is collected and often shared by a wide variety of organisations, from government sites such as those listed previously to commercial organisations and private citizens. It is up to you as a researcher to confirm the reliability of the data you are using.

Impressions	765
Total engagements	93
Detail expands	38
Media engagements	19
Likes	19
Profile clicks	10
Replies	2
Retweets	2
Hashtag clicks	2
Link clicks	1

FIGURE 1.9 An example of statistics available from Twitter to measure tweet impact and reach

Geographic Information System (GIS)

Referred to as GIS, the purpose of the Geographic Information System is to handle spatial or geographic data. Typically, GIS can overlay multiple sets of seemingly unrelated geocoded data (data that includes elevation/altitude, longitude and latitude) according to time of occurrence to show events in location. These are used to create maps that can show distribution of specific items or events. For example, by recording the home address of patients suffering from particular diseases, doctors can establish if there is a link with a particular location by plotting these on a GIS map.



Saruman-the-white [CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0)]

FIGURE 1.10 This example shows how GIS visualisations can combine map data with census data from the Australian Bureau of Statistics.

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

Referencing data sources

You must acknowledge all sources of data and information you use in your work. This means not only direct quotes, but also ideas, summarising and paraphrasing. There are numerous referencing styles available. All use a brief identifying **citation** in the body text (either in brackets or as a footnote) and a matching detailed reference in a reference list and/or a bibliography. A reference list only includes the works cited in the body text. A bibliography also includes other texts that were consulted and may be of interest to the reader but were not referred to in text. A bibliography can also be annotated. The sources of **intellectual property (IP)** that influence your work in Data Analytics should be acknowledged using the **American Psychological Association (APA)** referencing system. Note that APA style uses a reference list and not a bibliography.

This referencing style is not just applicable to this Outcome; it is also used for the first part of the SAT (Unit 3, Outcome 2). Referencing in this way ensures that:

- it is clear that you are not claiming another author's work as your own
- readers can find the original source to get more information
- **copyright** 'fair dealing' legalities are observed
- moral rights are observed.

If you fail to acknowledge someone else's intellectual property in your work, you may be accused of, and be guilty of, plagiarism. **Plagiarism**, which is when you pass someone else's original work off as your own, is very serious. It can have repercussions beyond embarrassment: you may find yourself in academic and legal difficulty. Your school or university may take disciplinary action. You may even face prosecution or civil legal action under copyright laws.

It is wise to start recording your sources as you find them, rather than at the end of the Outcome. Having to locate sources again to cite them may be difficult and time-consuming.

APA style guide

The American Psychological Association (APA) referencing system is widely used and is one of the most common styles for referencing. Citations appear within the text showing the author's last name, the year of the publication and the page number of a quote. These are usually separated by commas and are placed within parentheses following the text. For example:

> Despite the fact that fruit and vegetables seem, at first glance, similar enough in terms of dietary requirements, and tend to be used interchangeably, the topic of diet can be contentious. Fruit tends to have higher levels of fructose than vegetables and in general terms a fruit can be a vegetable, but a vegetable cannot be a fruit. (Redman, 2018, p. 231–6)

The corresponding source details appear in a reference list at the end of the work. Microsoft Word can easily be used to keep track of your sources. You just need to fill in the boxes and it can automatically create a reference list or bibliography for you. Tables 1.1 and 1.2 demonstrate how to set out simple citations in your reference list.

9780170440875

TABLE 1.1 Use of APA for a simple book author-date reference list citation. This is the most basic citation that would form the basis of your reference list.

APA	Family name, Initial(s).	(Year of publication).	*Title: Subtitle.*	Edition and volume if applicable (XX ed., Vol. XX).	Place of publication (state, city, suburb):	Publisher,
	Redman, N.	(2018).	*Minding My Own Diet: Fruit and Vegetables.*	(5th ed., Vol. 2).	South Melbourne, VIC:	Peanut Press.
	Redman, N. (2018). *Minding My Own Diet: Fruit and Vegetables*. (5th ed., Vol. 2). South Melbourne, VIC: Peanut Press.					

TABLE 1.2 Use of APA for a simple website reference list citation. You should check the legal section of any website if the web page and document you are using has no information about the author on it.

APA	Author or organisation who created the web page (if clearly identifiable).	(Year the web page was published, or most recently updated).	*Title (of the web page or document, should be in italics if it stands alone)*	[Format description if applicable].	Retrieved from URL.
	Redman, N.	(2018).	*Minding Your Own Diet*	[Blog post].	Retrieved from https://www.normanredman.com.
	Redman, N. (2018). *Minding Your Own Diet* [Blog post]. Retrieved from https://www.normanredman.com.				

Referencing tips

Quotations

- Do not correct errors in quotations. A quotation should be represented as it appears originally, even if it has errors. You can use '[*sic*]' to flag to a user that you have noticed an error in the quotation and that it has not come from you.
- Do not change the spelling in a quotation or a name to match Australian English. Preserve correct names and original spelling. For instance, if you quote the World Health Organization, or need to make note of the World Trade Center, you would not change their names to World Health Organisation and World Trade Centre.
- Use square brackets to add explanatory information or give context to a quotation; for example, Darcy's claim that 'during the [American Civil] war, infections killed more men than bullets ever did …'
- Use ellipses '…' to indicate when part of a quote has been omitted.

9780170440875

Reference lists

- You can use Microsoft Word to save a list of your sources in a document and produce a reference list or bibliography. On the 'References' ribbon, look for 'Bibliography' or 'Citations & Bibliography' (though this will depend on your version of Microsoft Word).
- Write 'n.d.' if no date is given for a website or document.
- If no page number is known, write 'n.p.' and provide an estimated page, paragraph number or nearby heading.
- Use unspaced en dashes rather than hyphens in your page reference spans. To insert en dash in Microsoft Word in Windows, hit CTRL and minus on the number pad, or ALT and 0150.
- References are presented alphabetically using the surname of the first author.
- In the reference list, use the hanging indent paragraph style.
- If the source has a digital object identifier (DOI), place it at the end of the reference. The DOI is a permanent, unique identifier for an object, such as a serial number for publications, and may look something like doi:10.1037/0022-006X.56.6.893. It never changes even if the document's location or name changes.

Data types

We have discussed a range of methods that organisations and individuals use to acquire data from users. In most cases, the data entry **input** occurs via an online form populated with boxes into which the data is entered. It is important to understand the **data types** that are used because ultimately any data collected will be stored in a database (see page 19) and the data itself will need to be sorted and queried.

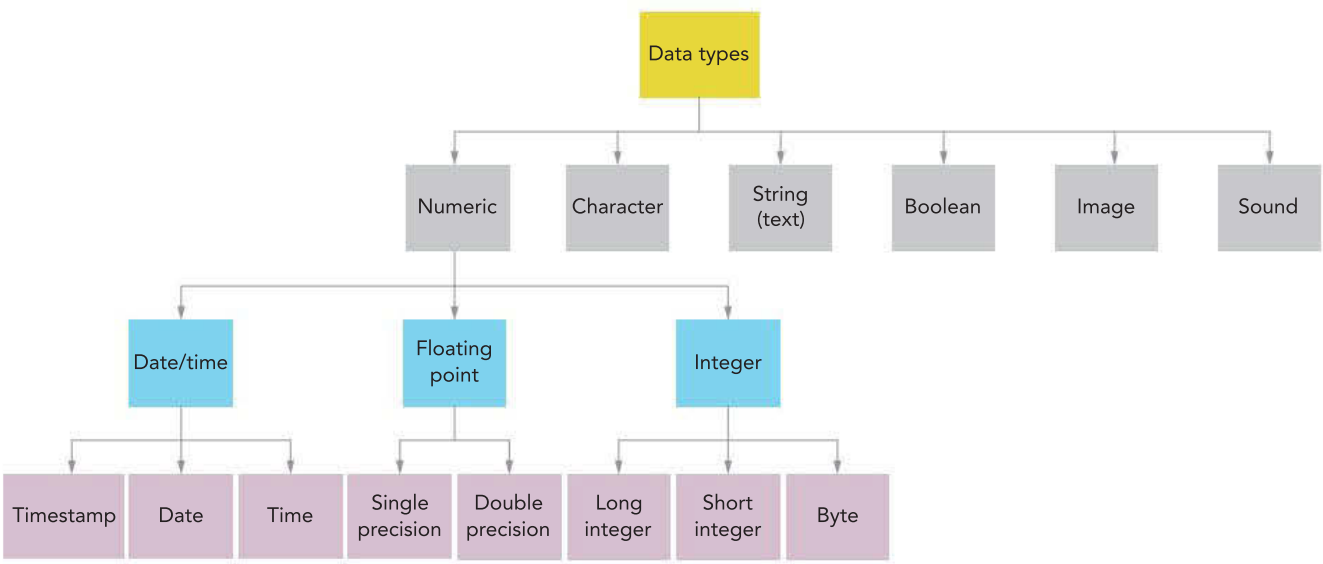


FIGURE 1.11 Diagram of different data types

9780170440875

Character

This is a text field that will only accept a single alphanumeric **character**. It is used where there are multiple options for a value, but they can be represented with a single character to make data entry easier and to save storage space. For example, the sizes of a wooden box that comes in small, medium and large sizes might be entered as 'S', 'M' or 'L' respectively. These options might be selected from a radio button group on a form.

Text – string

The majority of fields can and should be set as a **text data type**. Text may also be referred to as a 'string', particularly in connection with programming. This type of field holds a mix of characters (letters, numbers, special characters) – also referred to as alphanumeric – to a limit of 255. Names and addresses are considered text data. Postcodes and telephone numbers are formatted as text because they may contain spaces and are not intended to be used the same way as a numeric value. It is also more efficient to store the values as text rather than as a large numeric value.

Numeric – integer, floating point

Another key type of data format used in databases is numeric. These fields will only allow numbers to be entered. They are often used when the value is to be used in a calculation of some kind. For example, the quantity of an item purchased might need to be multiplied by its price to calculate a total amount payable. This calculation cannot be performed on a field formatted as text.

Numeric fields can also be categorised into different variations. **Integer** refers to whole numbers, including negative numbers. Where decimal numbers are required, such as when dealing with financial transactions or percentages, then the 'floating point' data type is used. Depending on the database package you use, the specific names of these **numeric data** types may vary, but their function will be the same.

Floating point is a term used mainly in software development.

Numeric – date

Strictly speaking, a date format is another variation of a numeric data type. The value used is normally based on the number of days since the 'zero' day built into the operating system or relational database management system (RDBMS). For example, day '1' might be displayed as 01 January 1900, while '44 134' would be displayed as 30 October 2020. Calculations can be performed on dates, which can be handy when comparing the difference between them. Dates can be formatted to show a combination of years, months, days, hours, minutes and seconds, depending on the needs of the user. In terms of time, they can also display 12- and 24-hour clocks.

Boolean

In cases where the data to be entered falls into the categories of Yes/No, True/False, 0/1, or even On/Off, the **Boolean data** type is used. This is often represented as a tick box on forms.

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

Image

The **image data** type is used for any graphical item such as photographs, diagrams, **charts** or illustrations. They would usually be stored as an image file (common choices include .jpg, .png and .gif) before being stored or manipulated. Images should be used sparingly in databases since they make the final files very large. For example, images of marine life may be included in a database to help users be certain they are looking at the correct record. Images are stored by computers in binary.

Sound

Sound data refers to any audio electronic recording. Common sound file types include .mp4, .wav, .aac, .flac, and the no longer supported .mp3. For example, a collector of bird calls may choose to record their collection using uncompressed audio files to preserve as much detail as possible, or they may decide to use compressed files to save on storage space and sacrifice some sound quality instead. Sound, like any image, is stored in binary.

MP3 was the most popular format for the distribution of audio files for many years. In 2017, the creators of the file type stopped supporting it, even though it is still the most popular audio file format used worldwide.

TABLE 1.3 The most common data types used in databases

Data type	Characteristics or uses	Examples
Text	Alphanumeric; up to 255 characters; searchable	Name, address, postcode and telephone number
Numeric	Numbers only (see below for different formats)	Any number that will be used in a calculation
Date	A variation of numbers, but formatted to represent a date and/or time	Any date; can be used in calculations
Character	A variation of a text field that can only hold a single alphanumeric character	Where space is to be saved, but there are more than two options to choose from
Boolean	Represents one of two states, such as True/False	Also represented as Yes/No and On/Off

THINK ABOUT DATA ANALYTICS 1.2

Find two different compressed and two different uncompressed audio formats and compare them. Which is better? Is file size more important or less important than sound quality?

THINK ABOUT DATA ANALYTICS 1.3

Take a screenshot of an online data-collection form and annotate each of the data input boxes to indicate its likely data type.

The following is additional information that you may find useful.

- Storing a date as date data type will allow databases, spreadsheets and programs to extract parts of the date (year, month and day).
- A field's data *type* is separate from its data *format*. A field of date type can be displayed as 2020/03/28 in a list on-screen and 28 March 2020 on a printed certificate. One half may be shown as 0.5 or 50%.
- A timestamp data type contains both a date and a time of day.
- Do not confuse a time of day with a time duration. A song's length should be entered as a number. For example, 80 seconds = 1 minute 20 seconds, not a time of day (such as 1:20 = twenty past one).
- None of the integer types can store fractional data.
- A value such as 1:20 is meaningless to a database. Similarly, 1 minute 20 seconds can only be stored as text, and the software will not be able to interpret the value.
- Small $1.40 or large $2.20 can be stored and displayed, but the numbers will never be accessible for calculation or individual display.

9780170440875

Data structures

To store your data, you should be familiar with data types and data structures. Data is categorised into types so it can be stored efficiently and processed effectively. These include the following data types described earlier in this chapter: numeric (integer and floating point), text (string), Boolean, character and date.

The basic data types are described starting page 16.

Spreadsheets

Spreadsheet applications such as Microsoft Excel, Google Sheets or Apple's Numbers essentially guess what data type to use, and sometimes guess incorrectly. If you enter 01/02 into a cell in Excel, it will be interpreted as a date. If you enter an apostrophe first, you can force Excel to treat it as text; that is, '01/02. Alternatively, you can change the cell's format to 'Text' to prevent the problem.

Only enter one piece of data (or one datum, the rarely used singular for data) of a single data type into a spreadsheet cell or data field. Do not store units, such as seconds or litres, in cells with data. Put the units in the column's heading, such as 'Capacity (litres)'.

More details around spreadsheet creation and use will be covered in Chapter 2.

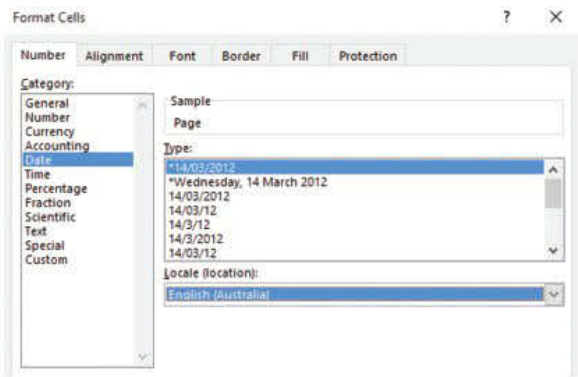


FIGURE 1.12 Formatting the appearance of dates in Microsoft Excel

Databases

Databases such as Microsoft Access, SQL and FileMaker have formal fields, records and table structures, while spreadsheets such as Microsoft Excel do not. In spreadsheets, cells do not need to be defined as containing a certain data type. You can format a cell's appearance, but its data type cannot be specified and imposed. The nature of data in a cell can remain ambiguous until a specific type of arithmetic or function is used to manipulate it, such as date, number or text.

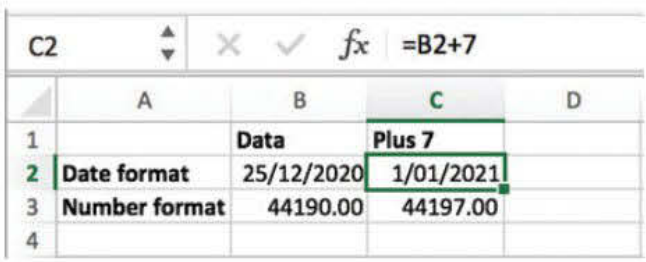


FIGURE 1.13 A spreadsheet handling the same data (column B) as dates and numbers

Remember: Having only one piece of data in a field is necessary to make all fields able to be searched and sorted.

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

Next steps

As you collect your data, take steps to protect your respondents and subjects, maintain **data integrity** and apply appropriate data types and structures, you should begin to think about relevant legal constraints. Earlier in this chapter, there was a discussion around the importance of properly acknowledging your sources. In later chapters, this will be incorporated with a discussion of legal requirements related to storage and communication of data and information.

Data integrity

For a computer to produce useful information, the data that is input into a database must have integrity. Whether you are storing, transmitting or archiving data, you must be sure that its integrity is maintained. Otherwise, the data may not be accessible when you need it.

In your Outcome, you will rely on quality data to answer a question. It is important for you to develop and apply knowledge of integrity of primary and secondary data to your Outcome. The following sections discuss factors that influence integrity of data, such as accuracy, authenticity, correctness, reasonableness, relevance and timeliness.

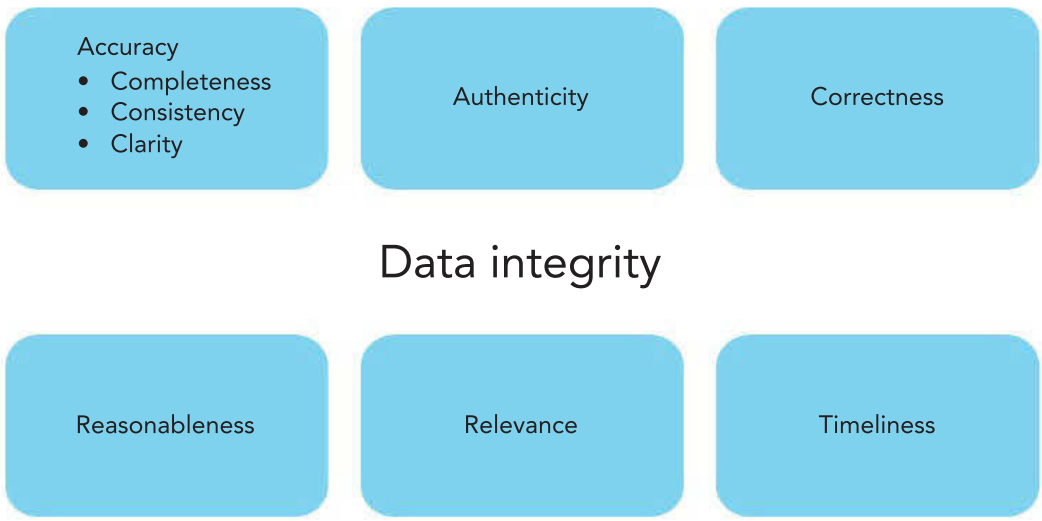


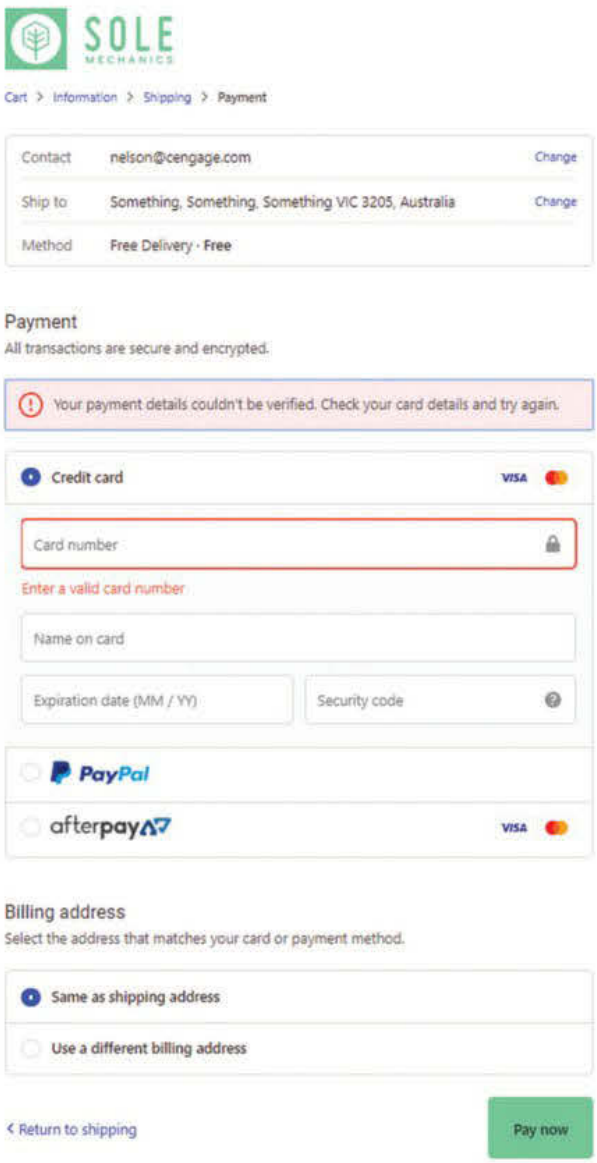
FIGURE 1.14 Factors of data integrity

Accuracy

Accuracy involves ensuring that the data collected is correct and does not contain any errors. When using primary methods to collect the data, **validation** may be able to be used to reduce the chances of incorrect data being entered. Data validation often involves restricting the data that can be entered into a particular field and by restricting what can be entered. The chances of incorrect data being input are therefore reduced.

Many online forms contain several validation methods that help reduce the chances of errors being input. Validation techniques include dropdown lists, radio buttons, predictive text, checkboxes and required fields.

9780170440875



Sole Mechanics – www.solemechanics.com.au

FIGURE 1.15 Online form containing validation techniques

1.4 THINK ABOUT DATA ANALYTICS

How many validation techniques are used in the online checkout form in Figure 1.15?

Completeness

Completeness means that your data set is just that: complete. You have data from all research participants on all variables at all relevant points in time and space. But completeness can be very difficult to achieve.

Consistency

Correct, unambiguous data can still cause a problem in a database if it is not consistent. Inconsistent data is unwelcome because it means the data is unreliable. This is why **consistency** is part of data integrity in this Outcome.

If you need to compare demographics as part of your Outcome, and some of your data came from subjects or respondents in inner Melbourne suburbs, you may make the mistake

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

of entering both 'St Kilda' and 'Saint Kilda' into your database. Both of these would be correct and they would be unambiguous. However, they would not be consistent. This is a problem because you would not be able to compare the data you have gathered properly because you would have data for two seemingly different suburbs instead of one suburb. If you looked directly at the data once it was entered into your database, you may notice that the entries are incorrect and fix them. However, it may seem insignificant enough that it does not catch your eye, even though it is there. Anomalies of this nature, or similar, prevent data being accurately aggregated and compared and increase the likelihood of generating inaccurate information.

Consistency is also a concern on a larger scale when it occurs between multiple data sources where conflicting versions of data appear in different places. If a 'true' value cannot be easily determined between multiple data sources, an entire data source loses integrity and becomes tainted.

Clarity

Clarity is about formatting data in an unambiguous manner to prevent misinterpretation. For example, you have entered the dates of birth of all participants into your database and they are the correct values. They are still not completely accurate, because it appears that some of the birth dates have been entered using the US date format (MM/DD/YYYY) and others using the Australian date format (DD/MM/YYYY). Worse still, some of the dates have both days *and* months under 12, making the actual interviewee subjects' dates of birth ambiguous:

- 04/11/2001 Grey, Jennifer
- 12/02/2004 Pierce, Jacob
- 06/09/2003 Martin, Gregory

The birth dates are inaccurate because you cannot tell what the correct values are. You need to be stricter when entering data. In doing this, you can apply consistency.

Authenticity

Digital documents are easy to fake and distort convincingly. It can be difficult to tell if an image, document, database or web page is genuine, a parody (for example, *The Onion*) or a deliberate attempt to lead, fool or defraud people, as with spam, phishing sites and links that trick visitors into clicking ads.

Authenticity relates to how genuine the data set is. Is the data genuine, original, accurate, reliable and trustworthy? The authenticity of primary data is easier to confirm since it has been collected first-hand, although precautions should be put into place to ensure it is accurate and reliable. Also, when collecting primary data, it is important that the **sample size** is not too small since this may lead to inaccurate results.

Characteristics of authentic data

Digital data can only be considered authentic (genuine) if it:

- comes from the author and/or source it claims to be from
- has not been deliberately corrupted
- is not faked or disguised as something else
- has not been changed without authorisation
- is what it claims to be and does not misrepresent itself
- does not aim to mislead or deceive by pretending to be anything else.

9780170440875

Correctness

Correctness means that the values stored for a given object must be correct. For example, if you create a database that includes a list of all the subjects participating in interviews, you would need to ensure that you enter each subject's details correctly into the database. The content would be incorrect if you keyed the wrong date of birth into the BirthDate field or misspelled a surname in the LastName field.

Data is also correct if it is a truthful representation of the real-world construct to which it refers. For example, the weather forecasts by the BOM are considered (usually) correct, in that they truthfully represent our *concept* of the weather in terms of predicted temperatures, rainfall, winds and so on.

Reasonableness

Data that is reasonable means that it falls within expected boundaries. **Reasonableness** refers to data that is believable and that checks are customarily carried out while data is being entered or inputted. Its purpose is to detect glaring errors or typographical mistakes. It cannot detect accuracy, and depends on users to decide what is and what is not reasonable. For example, a newborn baby's weight may reasonably be considered to be between 2 and 5 kilograms, but:

- a weight that is entered as 3.1 kilograms may or may not be correct
- there are weights outside this range that could also be accurate.

Relevance

People look for information that relates to a topic that interests them. **Relevance** measures how closely a resource, such as a book, database or web page, corresponds to people's desire for information.

Relevance is not always easy to measure. For example, it *is* obvious that income is relevant to spending habits. However, it is less clear whether age or gender are relevant to the development of schizophrenia. Data about schools in the United States could be relevant to Australian schools, but this is not absolute. Data about men may apply to women in some circumstances, but not in others. Assuming that data is relevant when it may not be can lead you to draw invalid conclusions.

Timeliness

Timeliness relates to the age of the data – how old is it? The data used should be relevant for the time period. For example, using Melbourne's population data from the 1990s to help plan the location for new primary schools would result in a plan that does not match the city's current needs. It is important that the data input into the information system is timely to match what is needed, and is not collected too early (or too late).

Timeliness can also relate to the information provided by the information system. If the information produced is not provided in a suitable timeframe, it would be useless. Imagine a school produces a daily bulletin outlining all the events that occur each day, but the bulletin is always published at the end of the school day, and not the start. The information is not being received in a timely fashion and, therefore, is useless to the organisation.

9780170440875

Data validation

Be careful not to confuse validation with testing or evaluation. These terms are often tested in examinations.

Validation checks that input data is *reasonable*. Validation does not and cannot check that inputs are accurate. How could validation tell whether a person is being honest when entering their age? Or that their name is correctly spelled? Validation, however, can detect problems when a person enters their age as 178 years, or 'fish', or nothing at all. You can perform validation manually (yourself) or allow software to do it for you electronically.

Manual validation

People can perform manual validation, especially proofreading for sense, clarity, relevance and appropriateness. In addition, unlike spreadsheets, people tend to notice when values entered would pass electronic validation checks but are inaccurate because they are ridiculous.

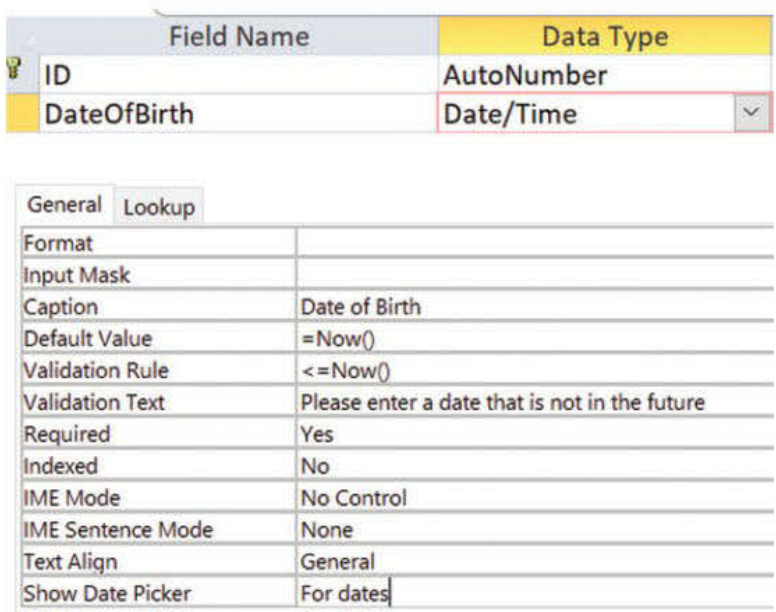
Similarly, Microsoft Word can find words that are not in its dictionary, but it cannot advise writers that a paragraph is boring or that an essay is pretentious, sexist and irrelevant.

Electronic validation

Computers are particularly good at conducting validation checks. These occur at the point of data entry or when importing. Any value or entry that does not meet the defined criteria either triggers an alert to the person entering the data or the data is simply rejected. There are several different ways that the data can be checked.

Electronic validation should be carried out in this order – there is no point in running a range check if the required numbers have not been entered. This will be covered in more detail in Chapter 2.

- **Existence checks** ensure that a value has been entered.
- **Type checks** ensure data is of the right type (for example, the age that has been entered is actually a number).
- **Range checks** ensure that data is within acceptable limits (for example, children enrolling in kindergarten must be 3–6 years old) or comes from a list of acceptable values (for example, small, medium or large).



Field Name	Data Type
ID	AutoNumber
DateOfBirth	Date/Time

General | Lookup

Format	
Input Mask	
Caption	Date of Birth
Default Value	=Now()
Validation Rule	<=Now()
Validation Text	Please enter a date that is not in the future
Required	Yes
Indexed	No
IME Mode	No Control
IME Sentence Mode	None
Text Align	General
Show Date Picker	For dates

FIGURE 1.16 Validation rules in MS Access are similar to those in FileMaker and MS Excel. Here, a date of birth field is made compulsory but not unique. The current date is set as a default value and prevents any dates in the future from being entered.

THINK ABOUT DATA ANALYTICS 1.5

In Unit 3, Outcome 1, what data will you mainly need to validate? How will this be achieved?

9780170440875

- **Input masks** are used to ensure that data is entered in the exact format required. This could be to check that a four-number postcode has been entered, or that a product code has the correct sequence of letters or numbers.
- **Dropdown lists** can be used to limit what data can be entered, ensuring that only acceptable pre-chosen values are used. This avoids issues around consistency as discussed earlier, ensuring that only one format is used, and that it is always spelled correctly (for example, only allowing Mr instead of also including mr, Mr., Mister)

Remember the following.

- Validation checks the reasonableness of data inputs.
- **Testing** checks the accuracy of information outputs.

In your Outcome, and later in your upcoming SAT, ensure that all of your data is thoroughly and appropriately validated.

Data visualisation

If you have difficulty interpreting your numeric data or want to make a point clear, you can use a **data visualisation**. Using (analogue) lines, shapes and colours to represent (digital) numbers can make it easier to interpret the data. Appropriate visualisation forms include **graphs**, charts, spatial relationships, maps, histograms and network diagrams. Each can be used to make trends and patterns more obvious.

You can represent different types of analogue data using dynamically changing physical indicators, such as strength, direction, duration or size of the signal. For example, a sound's pitch and volume, the position of a clock's hands, the height of liquid in a thermometer, or the length and angle of a line in a chart.

D3.js
Charted
Tableau Public
Infogram



Shutterstock.com/vasabii

FIGURE 1.17
Types of charts

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

Most versions of Microsoft Word have the same range of charts as Microsoft Excel. Click the Insert tab, and then click Chart in the Illustrations group.

Data visualisations can be created using different software tools, from spreadsheet applications like Microsoft Excel, Numbers and Google Sheets through to an array of specialised online applications including D3.js, Charted, Google Charts, Tableau Public, Infogram and many others. Keep checking online to see what you could possibly use since new tools are becoming available all the time.

The most important thing in any data visualisation is that it must be easy to see the main idea and trends in the data quickly and easily.

Charts

Charts are often used to visualise numeric data. There is a range of chart types, where each type is used for different purposes. A bar graph can be used to compare different items; a pie chart can be used to show each data item as a proportion of the population; line charts are useful for showing the trend in a data item over time; and histograms are useful for grouping data then showing the frequency of each group.

You can use spreadsheet software such as Microsoft Excel to produce various useful types of charts. Choose an appropriate chart type and follow its conventions.

Pie charts

A **pie chart** is divided into coloured or patterned 'slices' proportional to the percentage of the whole pie. Consider choosing a pie chart if you need to depict approximate proportional relationships (relative amounts) or compare part of a whole at a given point in time. The full circle represents 100%. The angle of each 'slice' is found by multiplying its percentage value by 360°. You should label the slices or use a legend (Figure 1.18).

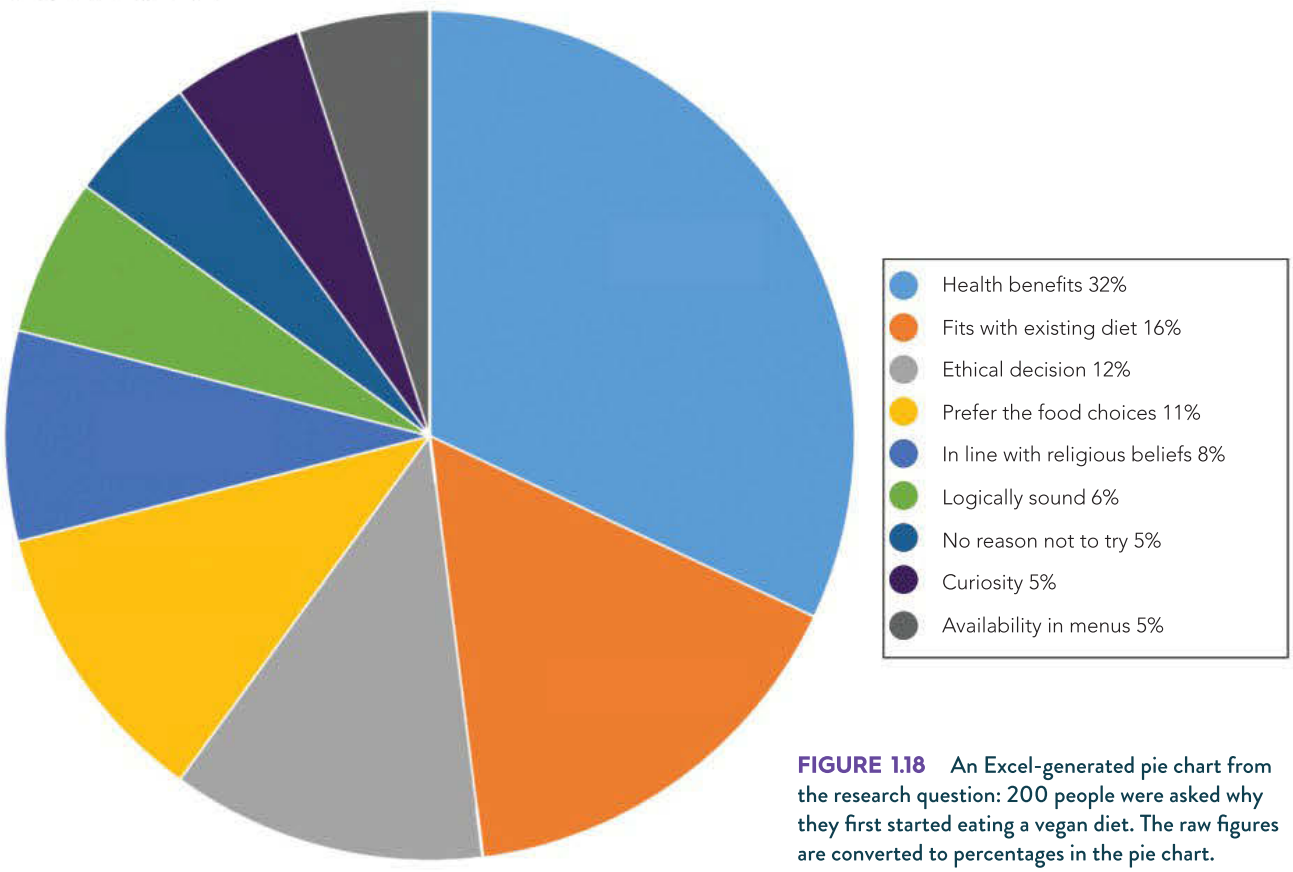


FIGURE 1.18 An Excel-generated pie chart from the research question: 200 people were asked why they first started eating a vegan diet. The raw figures are converted to percentages in the pie chart.

9780170440875

Histograms

Histograms display a series of values from a table. Differences in the height of the boxes make them visually striking and easy to interpret. The bars depict a range of related values so the bars touch. As an example, consider a series of age ranges and the number of people in each category (Figure 1.19).

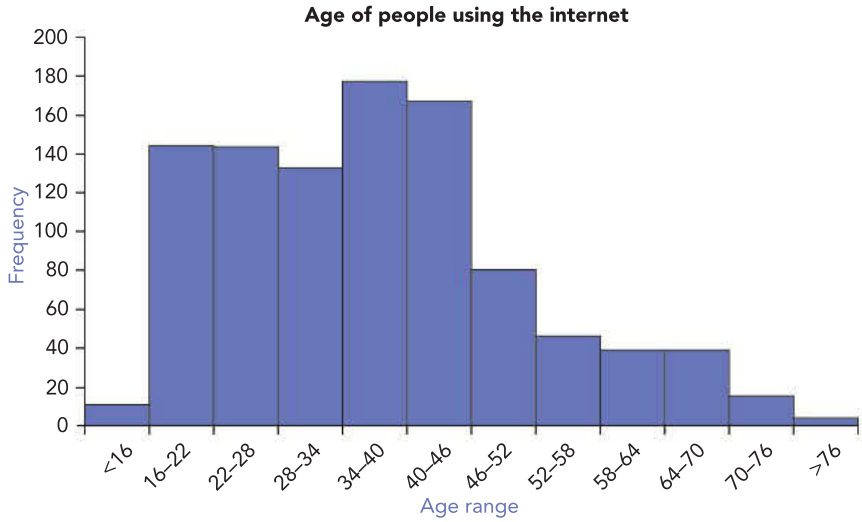


FIGURE 1.19
A histogram showing the age of internet users. Histograms are graphs with a continuous scale such as age groups.

Column graphs

Column graphs arrange data vertically and **bar graphs** arrange data horizontally (in rows). Unlike histograms, the categories are unrelated so the bars do not touch. An example of this could be a graph depicting favourite fruit (Figure 1.20). Both column and bar graphs are useful for presenting data changes over a period of time or for showing comparisons across different times. They enable visual comparisons easily so that differences are quickly recognised. You can use colours to distinguish between bars if representing different sets of data, but colour may be lost when printed.

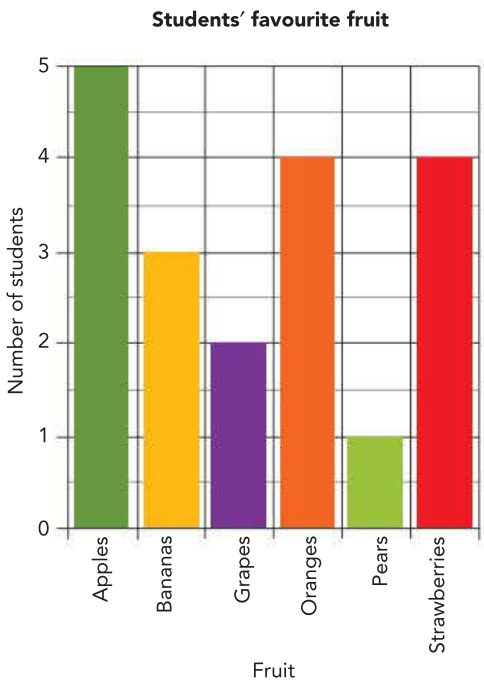


FIGURE 1.20 Column graphs show total numbers for data that is not part of a series.

Line graphs

Line graphs display continuous data over time, set against a common time. They are useful for showing trends in data at equal intervals, such as a graph depicting wheat exports over time (Figure 1.21, page 28). If more than one line is shown, you should use a different colour to distinguish each line.

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

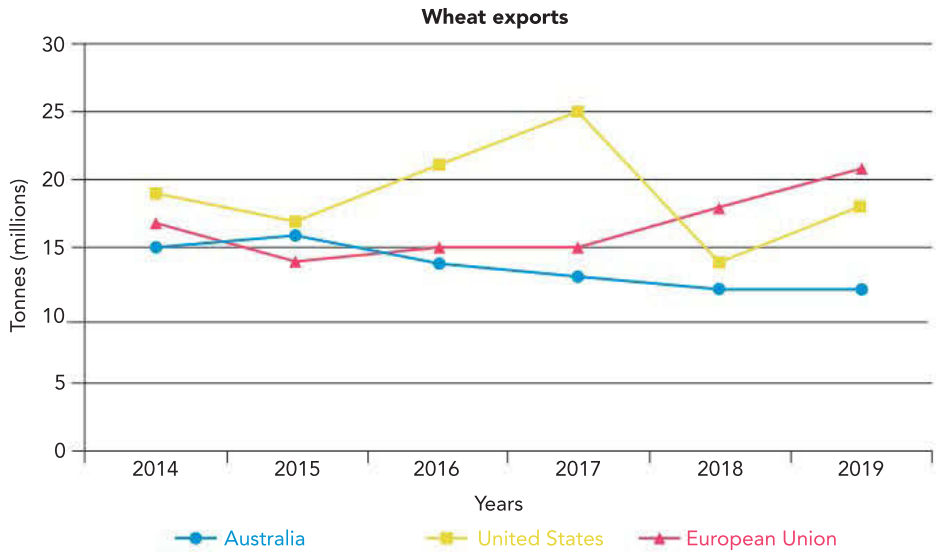


FIGURE 1.21 A line graph showing wheat exports across three different regions

Bubble charts

Bubble charts use circles to represent numbers – the bigger the circle, the larger the number. They can be used in a relative manner to compare the size of different cohorts or groups, or they can plot three variables onto a two-dimensional chart with an *x* and *y* axis. The size of the bubble (circle) represents the number of cases in each category. This can be seen in Figure 1.22, which compares income with life expectancy, while also representing the size of each country by the size of the bubble or circle.

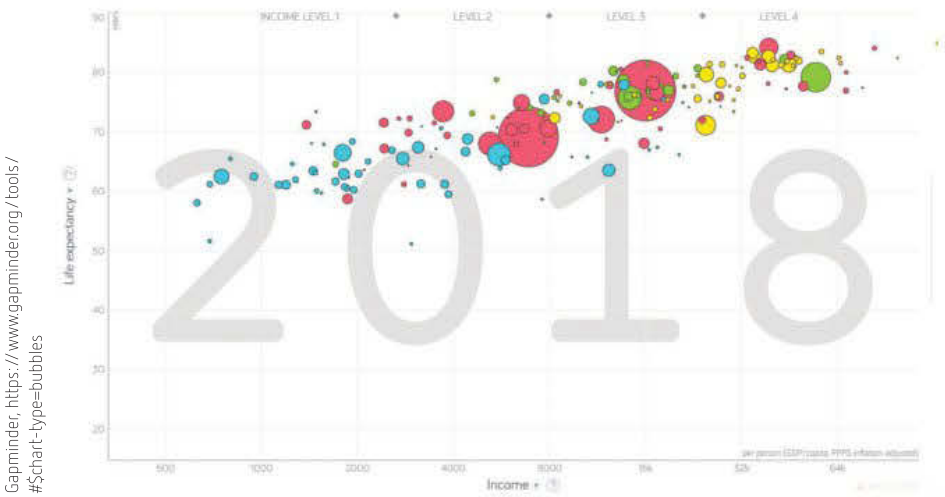


FIGURE 1.22 Bubble charts are able to visualise data such as populations with two other variables.

Spreadsheets can calculate lines of best fit for scattered data.

Scatter graphs

Scatter graphs show raw data points as dots. This is useful because data rarely lines up precisely, follows the mean exactly, or shows a perfect trend. You can add a **line of best fit** to your scatter graph (Figure 1.23, page 29), which will indicate the single trend that most accurately shows the general direction of the data's trend.

9780170440875

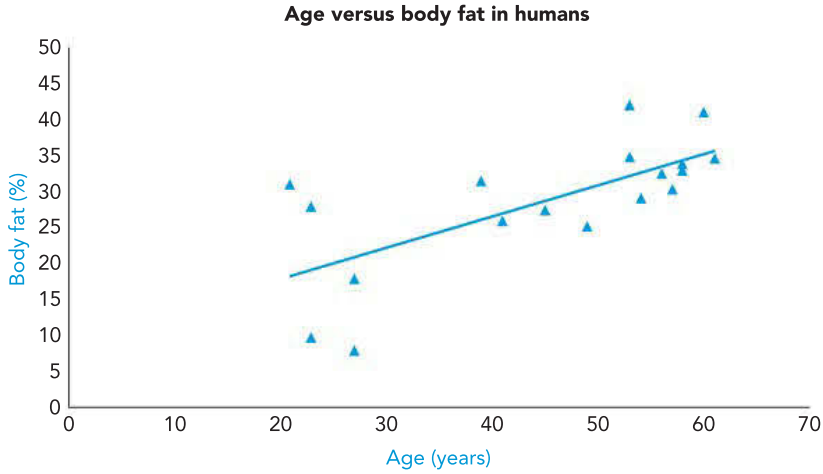


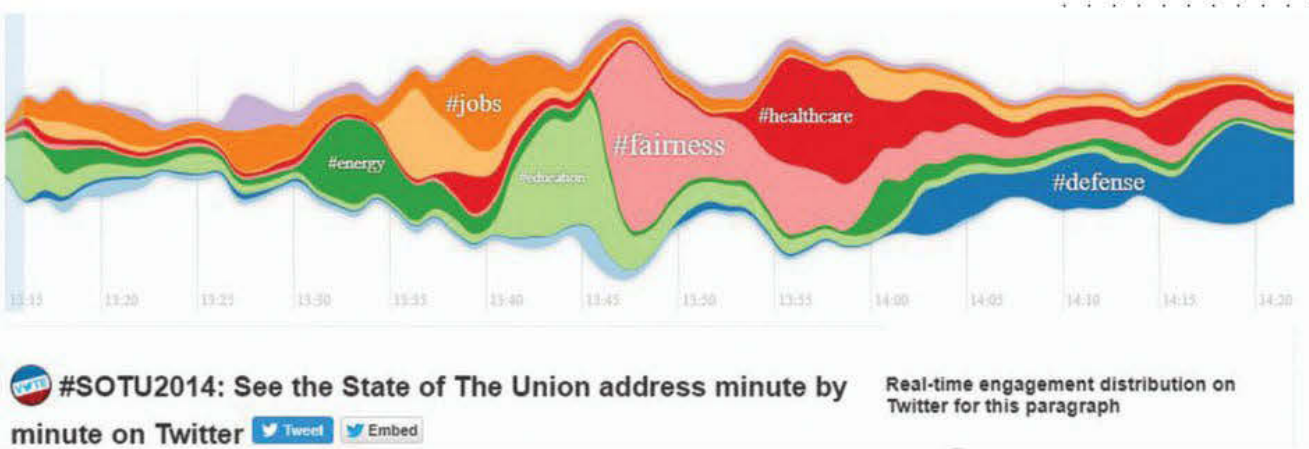
FIGURE 1.23 A scatter graph with a line of best fit: Age versus body fat in humans. The individual data points are hard to interpret. However, you can see the line of best fit markedly inclining upwards to the right, which indicates that body fat increases with age.

Interpretation of the line of best fit in Figure 1.23 indicates that body fat increases with age. A line of best fit also enables two other important statistical functions: interpolation and extrapolation. Interpolation involves calculating missing data using trends within the known data. There is no data on body fat for age 35 in Figure 1.23, but using the line of best fit, you can calculate that it would be approximately 25%. Extrapolation involves calculating missing data that is outside the limits of the known data. To find body fat levels for 10-year-olds or 70-year-olds, you could extend the line of best fit beyond the known data set and then read the anticipated body fat levels. Note that interpolation and extrapolation will only be accurate if the line of best fit follows a predictable path.

Microsoft has introduced sophisticated business intelligence tools into recent versions of MS Office, such as Excel, which offer data visualisation possibilities. You may want to investigate MS Power Map for visualising geographical data.

Stream graphs

Stream graphs can be used to show how data values change over time, such as the visualisation in Figure 1.24 created by Nicolas Belmonte to track people's reactions to then-president of the United States Barack Obama's State of the Union address in 2014.



Twitter/Nicolas Belmonte

FIGURE 1.24 Stream graph depicting real-time Twitter reactions to President Obama's State of the Union address in 2014

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

Time visualisations

Time visualisation represents a data item or data set over a period of time. Some time-based visualisations will show historical data, while others capture live data to provide real-time information. It is also possible to display the dimension of time by adding motion or animation to create a dynamic data representation.

A chart could be interactive, as seen with Figure 1.25, which shows several screenshots from a world wealth distribution chart on Gapminder.

The data could also be related to a timeline or time series. Timeline data may relate to individual items or events and shows the order in which the items or events occurred over a time period, while time series data may relate to the same data item and shows the variations or changes in the item over a time period.

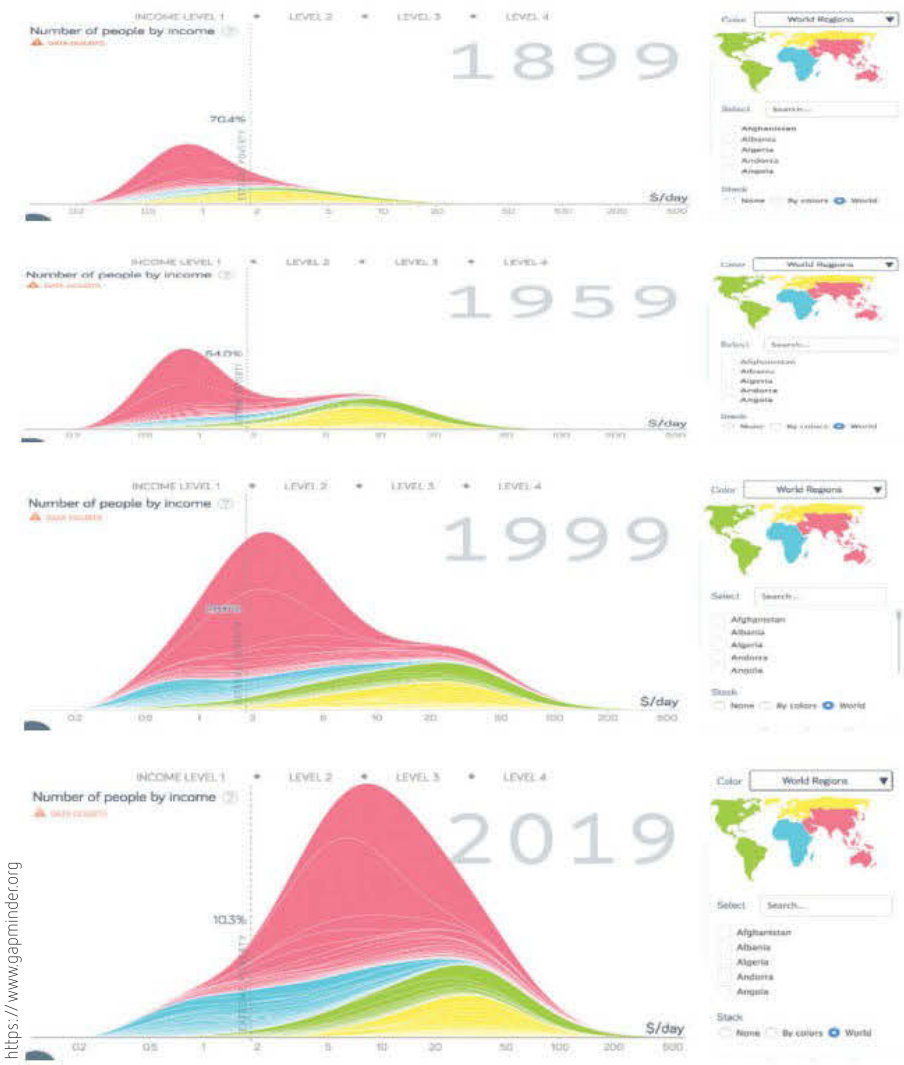


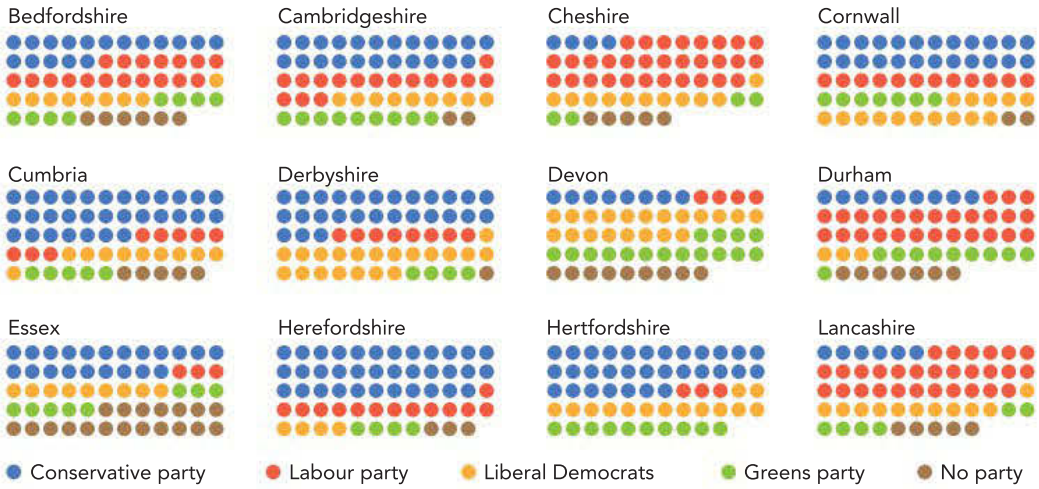
FIGURE 1.25
Screenshots from an interactive time series visualisation generated by OECD data by the Gapminder website (showing the shift in global income over the last 120 years)

Matrix visualisations

Matrix visualisations can be used to show the composition of individual items in the sample size. In this regard, they can be considered similar to pie charts. Matrix visualisations often divide the display area up into grids (similar to cells in a spreadsheet). Different sections of the display area are then used to represent the proportion of individual (or groups) of data items.

9780170440875

Matrix diagrams can be used to compare different data items, similar to a scatter plot. Again, the display area is divided in a grid or table format. Figure 1.26 shows several matrix visualisations with each visualisation representing a county in the United Kingdom. The colours represent the percentage of people who voted for a particular political party in an election.

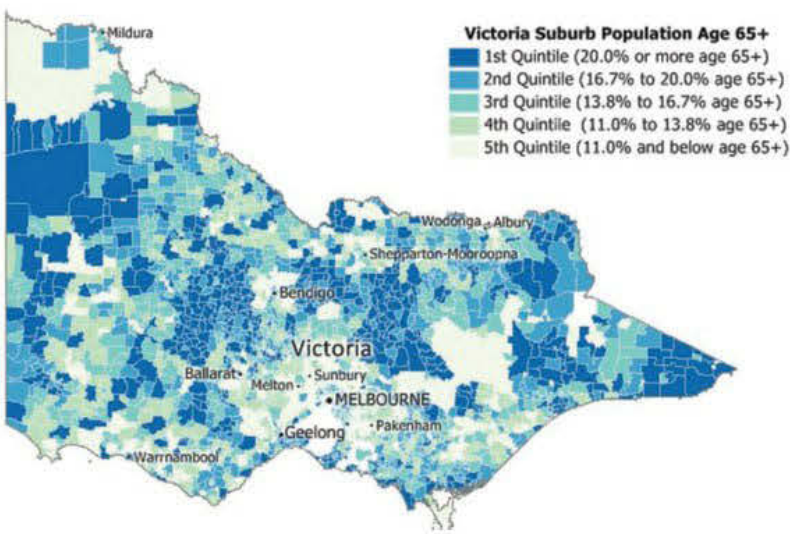


Christian Behrens, The Data Visualisation Catalogue

FIGURE 1.26 Matrix visualisation

GIS mapping

GIS data can be represented using maps to create visualisations of data as they relate to each other geographically. Heat maps are one example, showing where particular items or data are focused. They can be used to effectively display distributions of certain concentrations or distributions across an area. For example, Figure 1.27 shows the density of 65+ year old residents across the state of Victoria.



Produced with the Maptitude™ Mapping Software (MappingSoftware.com)

FIGURE 1.27 Map of Victoria showing where retirees reside

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

Conditional formatting

Conditional formatting uses software functions to change the appearance of data that matches a particular set of rules. This type of visualisation can be used on unprocessed data such as student results, rainfall totals or anything where the user's attention should be drawn to particular numbers or ranges.

Examine the sets of temperature data in figures 1.28, 1.29 and 1.30, displaying the recorded maximum daily temperatures in Melbourne for 2018. The data presented is identical in every way.

Figure 1.28 shows the raw data. It is difficult to get any ideas about how many hot or cold days there were or to identify any patterns. Figure 1.29 and Figure 1.30 (page 33), however, give you an immediate idea of the trends, highs and lows.

FIGURE 1.28
Spreadsheet showing Melbourne's maximum daily temperatures for 2018

2018 Maximum temperature (Degree C)												
Day	January	February	March	April	May	June	July	August	September	October	November	December
1	25.3	22.5	20.5	27.7	23.4	13.3	14.7	14.8	13.4	23.3	34.3	31.7
2	21.8	22.1	22.8	20.9	23.2	13.6	14.6	16.1	12.3	25.8	31.6	19.5
3	21	28.1	32.7	19.8	24.4	13.8	14.3	18.4	13.1	18.1	20.9	23.6
4	23.6	25.7	21	20.6	17.2	14.4	17.1	15.3	19.5	14.6	20.9	20.5
5	30.7	25.7	21.1	18.7	21.6	15	18.7	17.6	20.2	20.5	28.6	25.2
6	41.7	29.9	27.6	21.2	21.7	19.1	15.2	13.8	20.6	23.8	21.4	35.5
7	21.8	37.4	28.5	21.6	19.5	17.8	12.9	14.8	15.1	24.7	17.2	38
8	21.9	27.5	30.5	29.9	17.9	13.6	13.9	16	17.5	28.2	17	30.3
9	22.5	24.9	28.8	25.7	18.4	15.6	13.8	15.4	18.1	17.9	20	20.1
10	25.6	30.3	35.1	28.8	12.8	15.3	11.5	20	18	16.5	20.3	20.6
11	34.5	21.2	22.8	27.3	15.1	17.6	11.6	14.7	24.2	20.7	19.6	23.6
12	30	23.6	20.9	26.5	16.9	16.6	11.3	14.8	18.5	22.7	30.2	33.6
13	20.9	29	20.5	25.9	16.2	14.1	13.9	14.9	19.7	24.9	30.5	21.1
14	19.5	23.3	21.3	16.1	15.9	14.5	15.5	16.7	22.7	27.3	17.3	22.7
15	20.9	24.2	20.8	19.1	14.6	14	13.8	18.4	13.7	27.5	18.1	24.8
16	26.9	22.2	24.4	21	14.3	13.1	12.6	14.7	14	23.1	19	23.5
17	30.3	25.3	32	16.6	15	13.1	17.5	15.9	18.1	22.6	18.1	21.8
18	40	24	23.1	19.4	17.7	12	15.2	13.8	20.3	20.7	26.6	20.8
19	40.3	25.4	28	22.3	15.5	14.5	17.6	11.9	14.7	27.4	30.9	26.9
20	24.6	29.5	19.2	20.2	16.2	15.3	12.1	12.4	17.4	16.1	27.6	21.2
21	28.2	29.3	25.1	22.8	16.7	15.1	14	12.4	18.9	19.2	23.2	20.9
22	23.9	31	28.1	20.4	17	15	14.8	14	17.3	27	16.4	19.1
23	23.6	33.5	29.4	26.9	15.3	12.8	15.4	16	13.7	18.8	14.5	26.9
24	23.3	31.6	24.2	24.6	16.1	13.1	15.6	15.4	14.9	15.5	16.9	34.2
25	26.6	19.8	25.4	20.1	14.9	12.6	15.7	15.2	13.6	19.5	18.8	23.9
26	28.4	26.5	18.5	17.5	20.5	14.1	15.1	14.7	17.6	18.3	18.6	23.8
27	33.2	31.3	22.6	17.8	20.7	11.2	19.5	12.9	24.2	19.6	20	37.4
28	38.1	30.9	30.3	18.3	20.5	10.6	18.6	11.2	15.2	15.9	20.8	36.4
29	33.5		21.8	18	17.9	12.5	14.6	12.5	14	17	18.7	23.9
30	21		23.3	19.6	15.1	14.2	12.6	16.7	16.5	25.2	20.8	22.7
31	19.1		22.4		14.8		16.2	16		19.5		23.5

Bureau of Meteorology

Be very careful if using automatically generated colour scales. The conventions usually used (red = bad/low, green = good/high) do not always work; here, the convention for temperatures is reversed so that red indicates high temperatures.

FIGURE 1.29
Spreadsheet showing Melbourne's maximum daily temperatures for 2018 with colour scale applied. Note that it is much easier to get an idea of the distribution of temperatures.

2018 Maximum temperature (Degree C)												
Day	January	February	March	April	May	June	July	August	September	October	November	December
1	25.3	22.5	20.5	27.7	23.4	13.3	14.7	14.8	13.4	23.3	34.3	31.7
2	21.8	22.1	22.8	20.9	23.2	13.6	14.6	16.1	12.3	25.8	31.6	19.5
3	21	28.1	32.7	19.8	24.4	13.8	14.3	18.4	13.1	18.1	20.9	23.6
4	23.6	25.7	21	20.6	17.2	14.4	17.1	15.3	19.5	14.6	20.9	20.5
5	30.7	25.7	21.1	18.7	21.6	15	18.7	17.6	20.2	20.5	28.6	25.2
6	41.7	29.9	27.6	21.2	21.7	19.1	15.2	13.8	20.6	23.8	21.4	35.5
7	21.8	37.4	28.5	21.6	19.5	17.8	12.9	14.8	15.1	24.7	17.2	38
8	21.9	27.5	30.5	29.9	17.9	13.6	13.9	16	17.5	28.2	17	30.3
9	22.5	24.9	28.8	25.7	18.4	15.6	13.8	15.4	18.1	17.9	20	20.1
10	25.6	30.3	35.1	28.8	12.8	15.3	11.5	20	18	16.5	20.3	20.6
11	34.5	21.2	22.8	27.3	15.1	17.6	11.6	14.7	24.2	20.7	19.6	23.6
12	30	23.6	20.9	26.5	16.9	16.6	11.3	14.8	18.5	22.7	30.2	33.6
13	20.9	29	20.5	25.9	16.2	14.1	13.9	14.9	19.7	24.9	30.5	21.1
14	19.5	23.3	21.3	16.1	15.9	14.5	15.5	16.7	22.7	27.3	17.3	22.7
15	20.9	24.2	20.8	19.1	14.6	14	13.8	18.4	13.7	27.5	18.1	24.8
16	26.9	22.2	24.4	21	14.3	13.1	12.6	14.7	14	23.1	19	23.5
17	30.3	25.3	32	16.6	15	13.1	17.5	15.9	18.1	22.6	18.1	21.8
18	40	24	23.1	19.4	17.7	12	15.2	13.8	20.3	20.7	26.6	20.8
19	40.3	25.4	28	22.3	15.5	14.5	17.6	11.9	14.7	27.4	30.9	26.9
20	24.6	29.5	19.2	20.2	16.2	15.3	12.1	12.4	17.4	16.1	27.6	21.2
21	28.2	29.3	25.1	22.8	16.7	15.1	14	12.4	18.9	19.2	23.2	20.9
22	23.9	31	28.1	20.4	17	15	14.8	14	17.3	27	16.4	19.1
23	23.6	33.5	29.4	26.9	15.3	12.8	15.4	16	13.7	18.8	14.5	26.9
24	23.3	31.6	24.2	24.6	16.1	13.1	15.6	15.4	14.9	15.5	16.9	34.2
25	26.6	19.8	25.4	20.1	14.9	12.6	15.7	15.2	13.6	19.5	18.8	23.9
26	28.4	26.5	18.5	17.5	20.5	14.1	15.1	14.7	17.6	18.3	18.6	23.8
27	33.2	31.3	22.6	17.8	20.7	11.2	19.5	12.9	24.2	19.6	20	37.4
28	38.1	30.9	30.3	18.3	20.5	10.6	18.6	11.2	15.2	15.9	20.8	36.4
29	33.5		21.8	18	17.9	12.5	14.6	12.5	14	17	18.7	23.9
30	21		23.3	19.6	15.1	14.2	12.6	16.7	16.5	25.2	20.8	22.7
31	19.1		22.4		14.8		16.2	16		19.5		23.5

Bureau of Meteorology

9780170440875

2018 Maximum temperature (Degree C)

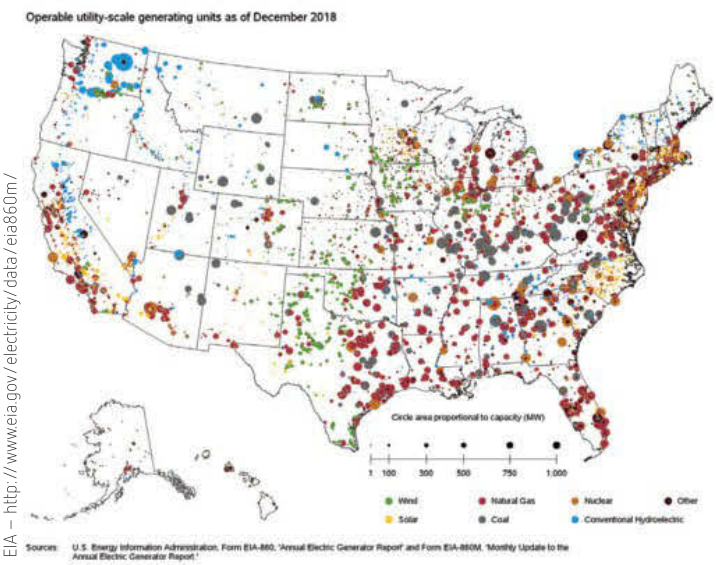
Day	January	February	March	April	May	June	July	August	September	October	November	December
1	25.3	22.5	20.5	27.7	23.4	13.3	14.7	14.8	13.4	23.3	34.3	31.7
2	21.8	22.1	22.8	20.9	23.2	13.6	14.6	16.1	12.3	25.8	31.6	19.5
3	21	28.1	32.7	19.8	24.4	13.8	14.3	18.4	13.1	18.1	20.9	23.6
4	23.6	25.7	21	20.6	17.2	14.4	17.1	15.3	19.5	14.6	20.9	20.5
5	30.7	25.7	21.1	18.7	21.6	15	18.7	17.6	20.2	20.5	28.6	25.2
6	41.7	29.9	27.6	21.2	21.7	19.1	15.2	13.8	20.6	23.8	21.4	35.5
7	21.8	37.4	28.5	21.6	19.5	17.8	12.9	14.8	15.1	24.7	17.2	38
8	21.9	27.5	30.5	29.9	17.9	13.6	13.9	16	17.5	28.2	17	30.3
9	22.5	24.9	28.8	25.7	18.4	15.6	13.8	15.4	18.1	17.9	20	20.1
10	25.6	30.3	35.1	28.8	12.8	15.3	11.5	20	18	16.5	20.3	20.6
11	34.5	21.2	22.8	27.3	15.1	17.6	11.6	14.7	24.2	20.7	19.6	23.6
12	30	23.6	20.9	26.5	16.9	16.6	11.3	14.8	18.5	22.7	30.2	33.6
13	20.9	29	20.5	25.9	16.2	14.1	13.9	14.9	19.7	24.9	30.5	21.1
14	19.5	23.3	21.3	16.1	15.9	14.5	15.5	16.7	22.7	27.3	17.3	22.7
15	20.9	24.2	20.8	19.1	14.6	14	13.8	18.4	13.7	27.5	18.1	24.8
16	26.9	22.2	24.4	21	14.3	13.1	12.6	14.7	14	23.1	19	23.5
17	30.3	25.3	32	16.6	15	13.1	17.5	15.9	18.1	22.6	18.1	21.8
18	40	24	23.1	19.4	17.7	12	15.2	13.8	20.3	20.7	26.6	20.8
19	40.3	25.4	28	22.3	15.5	14.5	17.6	11.9	14.7	27.4	30.9	26.9
20	24.6	29.5	19.2	20.2	16.2	15.3	12.1	12.4	17.4	16.1	27.6	21.2
21	28.2	29.3	25.1	22.8	16.7	15.1	14	12.4	18.9	19.2	23.2	20.9
22	23.9	31	28.1	20.4	17	15	14.8	14	17.3	27	16.4	19.1
23	23.6	33.5	29.4	26.9	15.3	12.8	15.4	16	13.7	18.8	14.5	26.9
24	23.3	31.6	24.2	24.6	16.1	13.1	15.6	15.4	14.9	15.5	16.9	34.2
25	26.6	19.8	25.4	20.1	14.9	12.6	15.7	15.2	13.6	19.5	18.8	23.9
26	28.4	26.5	18.5	17.5	20.5	14.1	15.1	14.7	17.6	18.3	18.6	23.8
27	33.2	31.3	22.6	17.8	20.7	11.2	19.5	12.9	24.2	19.6	20	37.4
28	38.1	30.9	30.3	18.3	20.5	10.6	18.6	11.2	15.2	15.9	20.8	36.4
29	33.5		21.8	18	17.9	12.5	14.6	12.5	14	17	18.7	23.9
30	21		23.3	19.6	15.1	14.2	12.6	16.7	16.5	25.2	20.8	22.7
31	19.1		22.4		14.8		16.2	16		19.5		23.5

Bureau of Meteorology

FIGURE 1.30
Spreadsheet showing Melbourne's maximum daily temperatures for 2018 with customised conditional formatting using three rules: yellow for temperatures less than or equal to 15, pink for 35 and above, red for 40 and above

Map-based visualisations

A popular method to display geographical data is by using **map-based visualisations**. These types of visualisation are often called geospatial visualisations. Geospatial data is related to the geographical location covered. Data could be related to population, roads, rivers, climate, mobile phone towers or any other characteristic of the area. Many geospatial visualisations are dynamic and allow the user to zoom in or out or navigate over an area. Geospatial visualisations are becoming more popular because they are a powerful tool that allows the data to be brought to life through visualisation. Because a range of data can be overlayed with a geographical location, the uses of these types of visualisations are enormous. Common uses have been for agricultural, environmental, mining and urban planning purposes, but the list is endless.



EIA – http://www.eia.gov/electricity/data/eia860m/

FIGURE 1.31
Geospatial visualisation

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

Hierarchy visualisations

Hierarchy visualisations show the relationships and structure between data items. In a hierarchy, data items are represented as being above, equal or below other items in the data set. Types of hierarchy visualisations can include organisational, structure and tree charts and mind maps.

They can be used to show the relationship between items in a data set or they can illustrate the breaking down of items into smaller components. Hierarchy visualisations are also useful for representing non-numerical data types. Figure 1.32 shows a word or mind map about computer networks. The subject networks have been broken down into sections and then the components of each section branch out from the centre.

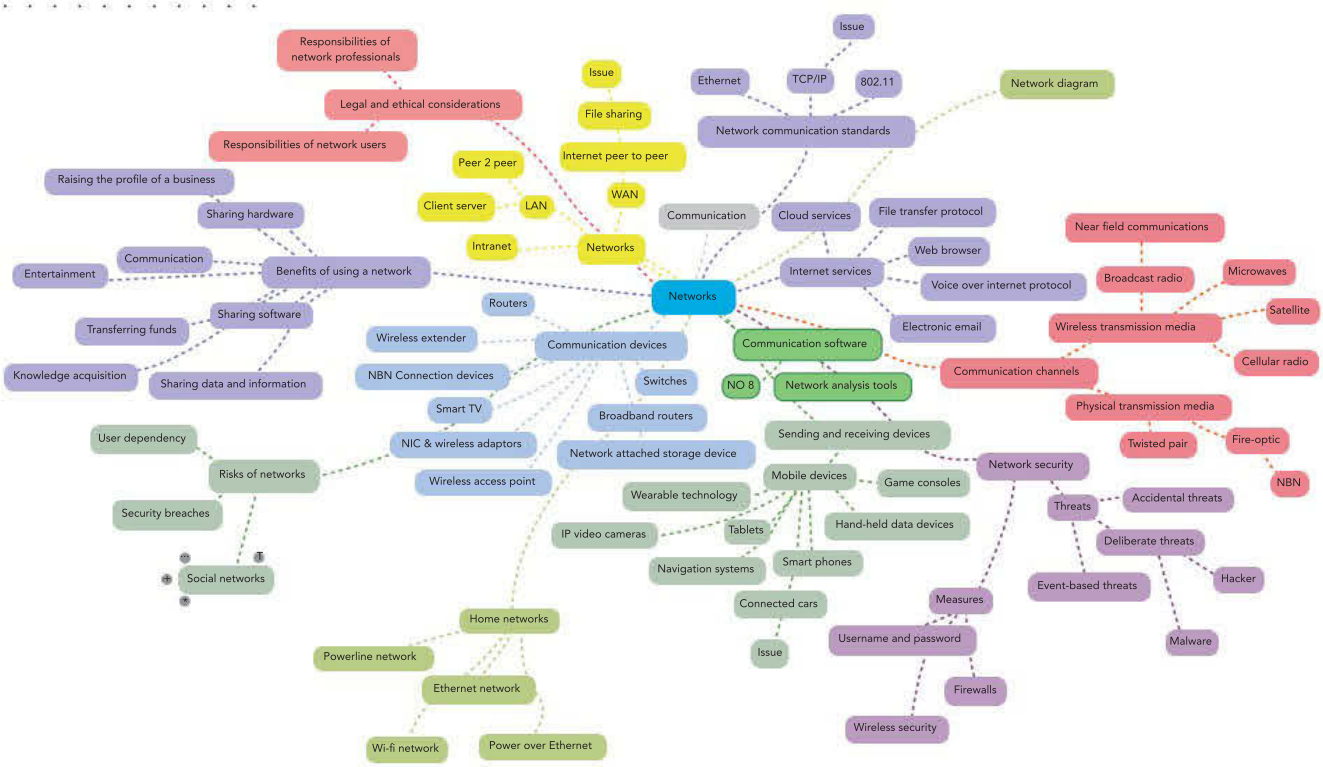
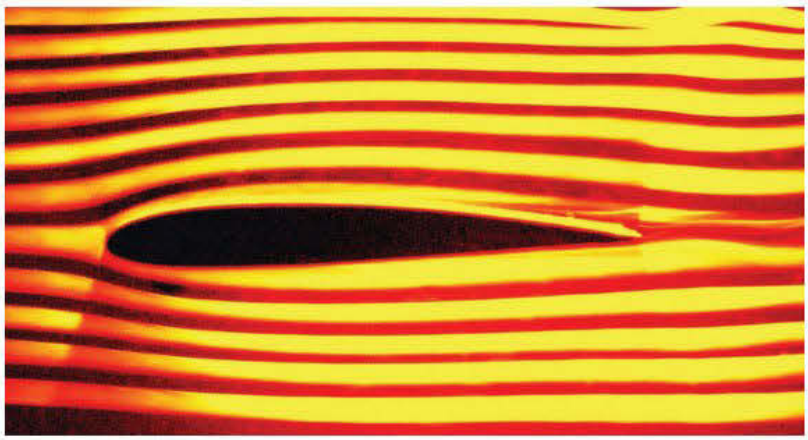


FIGURE 1.32 Hierarchy visualisation of notes about computer networks

Flow visualisations

Flow visualisations involve representing data that illustrates the flow pattern of a data item or items. This could be the pattern of customer movements through a supermarket or the series of pages a user would visit on a website to complete a transaction.

Flow visualisations are also used for scientific purposes to visualise the flow patterns of objects that are normally invisible, including air and water. Figure 1.33 represents the effect of an aircraft wing on the airflow passing the wing. The data for this was collected during testing using a wind tunnel and the data converted to a visualisation.



Science Photo Library/Dr Gary Settles

FIGURE 1.33 Airflow visualisation

Word clouds

Online tools to generate visualisations based on the frequency of words in a particular piece of writing are a popular way to help users get the main ideas being conveyed. It is important to check these carefully when they are created, since it is easy to skew results if the data is not cleansed properly. Small words such as *the*, *a*, *an*, *if* and so forth should be ignored. Words with the same root should also be combined; for example, *play*, *played*, *plays*, *playing* should not (generally) be used as separate words since they convey the same idea.

Websites such as Wordle and Tagxedo are among the many free tools online to create visualisations of text.

There will be detailed discussion about data cleansing in later chapters.



Tagxedo (http://www.tagxedo.com)

FIGURE 1.34 Visualisation of unedited text from Act 2 of Shakespeare's *Romeo and Juliet* using Tagxedo

Word clouds are excellent ways to deal with large amounts of qualitative data such as survey responses or research articles. Figure 1.34 shows a visualisation of *Romeo and Juliet* by Shakespeare using Tagxedo's website. A quick look is enough to work out that Act 2 of the play deals with positive emotions, since the larger words are the ones that appear more frequently in the text. In this case, the characters that speak more frequently are the largest words, then the rest of them are in decreasing size.

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

Infographics

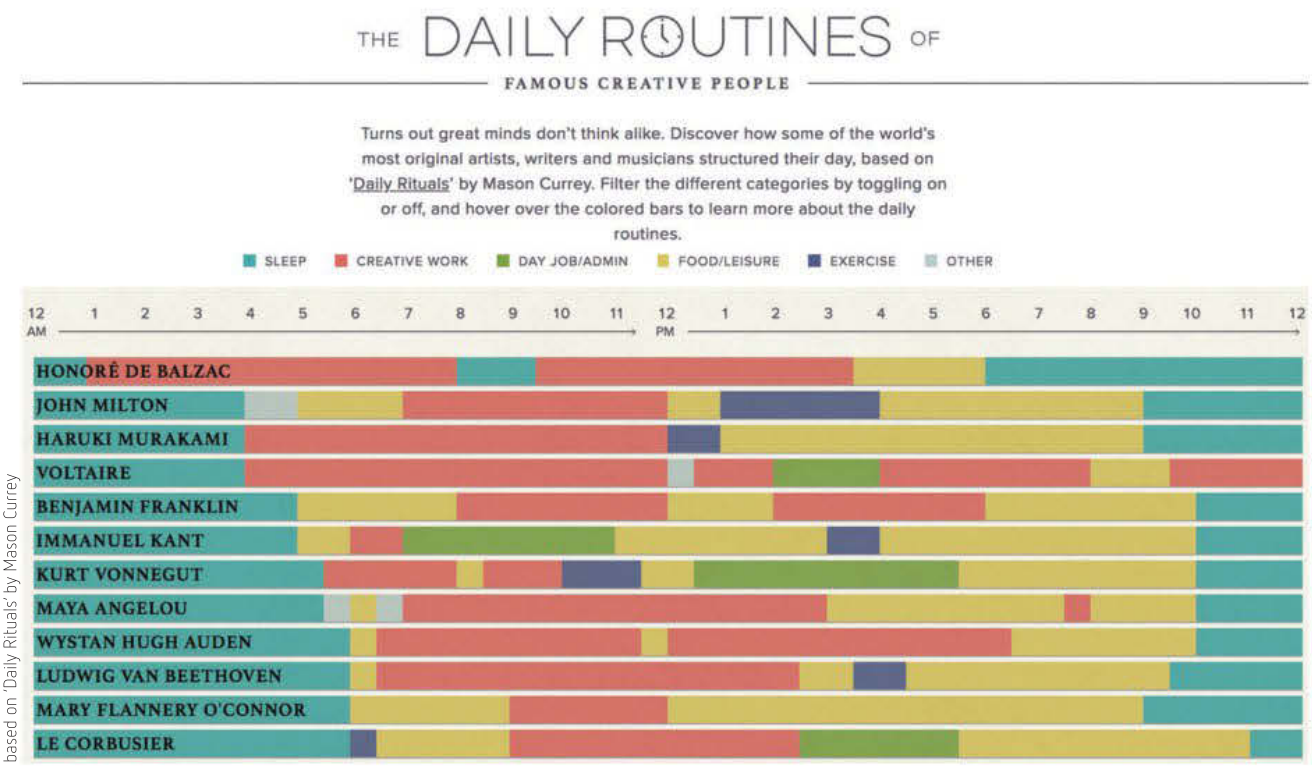
Infographics are becoming increasingly popular, due to their visual appeal as well as their ability to display large amounts of data in easily understood formats.

Infographics should not be text-laden. Lengthy text can take longer to digest, while there is more immediacy with images. By using a variety of tools such as shapes, charts, icons and diagrams, you can assist readers with visualising the data. Given that the purpose of an infographic is to provide information in a visual format, using visual cues will assist the reader with interpreting the infographic.

Infographics have no fixed formats or conventions, other than being highly visual and utilising a range of charts, diagrams or images to communicate their message to their intended users both efficiently (quickly and easily) and effectively (information successfully and correctly absorbed).

Figure 1.35 is an extract of an interactive online infographic that summarises the daily routines of famous creative people. Go to the weblink and scroll over sections that provide more detail and links to evidence.

The daily routines of famous creative people



'The daily routines of famous creative people', https://podio.com/site/creative-routines, based on 'Daily Rituals' by Mason Currey

FIGURE 1.35 Partial screenshot of an interactive infographic that illustrates how creative minds in history have spent their time. Based on research by Mason Currey.

9780170440875

Designing solutions

Design briefs

Once an information problem has been identified, the problem-solving methodology (PSM) outlines the first priority as being **analysis**, where the problem is broken down and its components identified. Once this is completed, it is usual to create a **design brief**. This is a documented account of everything that the proposed solution should do. This document is used to inform the next stage of the PSM – **design** – where designers will ensure that their work meets the requirements identified during analysis. There are several sections to a design brief: ***solution requirements***, *constraints* and *scope*.

Solution requirements

This section of the design brief lists the things that the solution (in this Outcome's case, a data visualisation) should *do* and should *be*. You may decide that your visualisation should address the economic impact of global warming for a teenage audience, and that it should be easy to understand and be visually appealing.

Constraints

Constraints address any limitations on your project or software solution. These are not usually decided by you. Constraints may include:

- economic (having limited time for a student or a set budget in an organisation)
- technical (not having a fast enough processor to create animated visualisations)
- availability of equipment (not having access to the hardware or software you need)
- security (not being able to store sensitive data safely)
- social (end-users not having a good experience using your solution)
- legal (not having permission to use someone else's data or ideas, or privacy issues)
- usability (the solution being too complex or incomplete to function properly).

Scope

Scope is where the boundaries of the problem are defined. They need to list clearly what will and will not be included in the proposed solution. While it might be beneficial to include all data from every study undertaken anywhere in the world, this is not practical in most cases. You might decide to focus on data only gathered in Australia between the years 2016 and 2019. You may choose to only use data that is free to use, or you may decide that only primary data collected by you or a trusted organisation will be used. It is important to clearly define what is out of scope as well, so that time and effort are not wasted gathering or analysing data that will not be used.

Design principles

Design principles are widely accepted approaches to creating designs for a variety of software solutions. This subject requires you to learn about the specific types of principles that need to be considered when planning new solutions. There are two categories of design principles – those relating to function (how it works), and those relating to appearance (how it looks).

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

Functional design principles

Functional design principles describe what needs to be considered when thinking about the non-visual aspects of your solution. The things you need to think about when doing this include:

- robustness – the solution should be able to run and not cause things like computer crashes
- flexibility – the solution should be able to work on different browsers and devices
- ease of use – the solution should be easy to use
- accessibility – the solution should be straightforward to navigate.

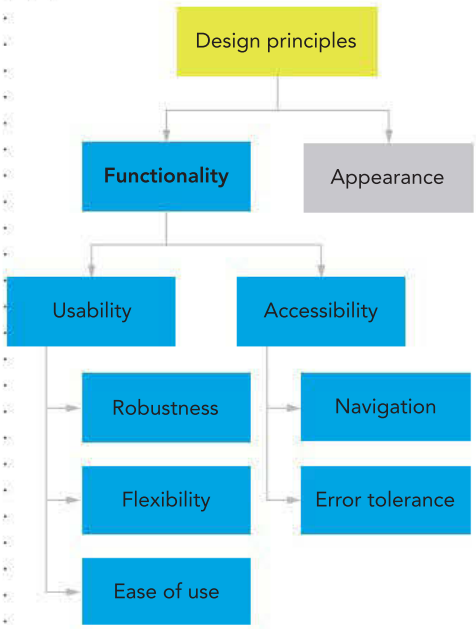


FIGURE 1.36 Functional design principles

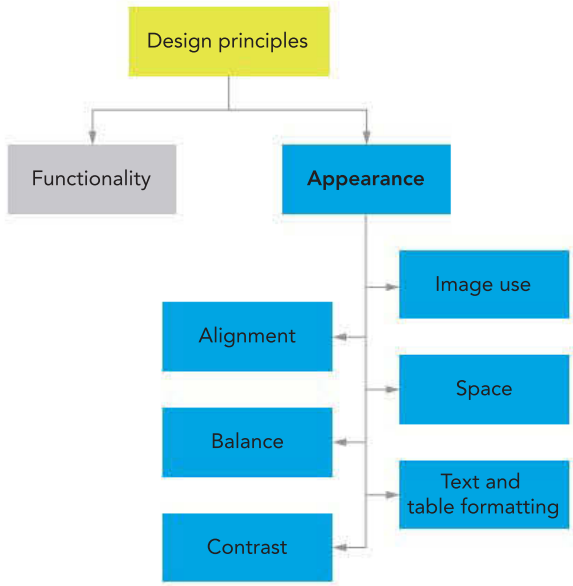


FIGURE 1.37 Appearance design principles

Appearance design principles

The design principles concerned with appearance are meant to make the designer think about the visual impact their designs will have. Items that must be considered when deciding how to arrange elements are:

- text formatting – includes font choice, alignment, colour, style (bold, italic), size
- table formatting – includes cell fill colours, header rows, borders, alignment
- alignment – horizontal (left, centre, right) and vertical (top, middle, bottom)
- contrast – high contrast is desirable – aim for very dark text on a very light background or vice versa
- image use – choosing images that are appropriate, relevant and do not distract from the message being communicated
- space – avoiding having large areas of empty space (use some space around important information to draw attention)

9780170440875

- balance – considering where elements are positioned – if there is a large element on one side, an equally weighted one could be on the other side; alternatively, two or three smaller items could serve the same purpose.

Formats and conventions

It is important to consider which charts, fonts, colours and other properties will be used when creating visualisations for other people to use. There are a few things you need to consider before making these choices.

It is important to consider the **intended users** – those who will be using your completed solutions. If the users belong to a specific group of people, such as university science students, then you can likely assume some previous understanding. If the audience is primary school aged or does not speak English as their first language, then that too will impact the final product that you make. If an audience is specified, you may have to do some research to discover if they have any particular needs that you should address. Cultural awareness may be needed in some cases to avoid using colours associated with mourning or using images of deceased persons. Careful research into your audience's preferences will help produce a successful product.

Consider what it is that you are trying to convey through your visualisations. **Clarity of message** addresses whether or not someone in your target audience of intended users understands what the central idea or ideas are. If they can correctly interpret your visualisation, then it can be considered successful.

Formats

Format refers to any choices made regarding how something, such as a data visualisation, should look. It refers to which tool(s) should be selected, such as a chart, table or dynamic online charting tool, as well as any choices made regarding appearance, such as the number of decimal places used, alignment, or the use of symbols like dollar signs.

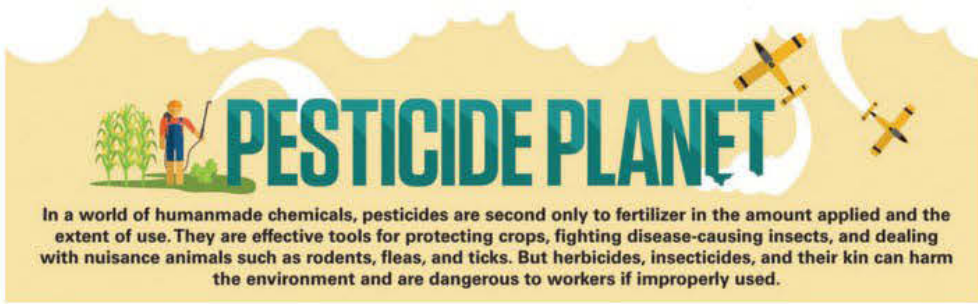
Conventions

In addition to the basic design principles you should follow when creating your graphic solution, there are several useful formats and **conventions** suitable for graphic solutions, such as titles, text styles, shapes, lines and arrows, sources of data and legend, and colours and contrasts.

Titles

In the simplest of terms, adding a title to a document makes it a dominant element. Titles are generally styled as headings, with type that is bold and larger than the body text or subheadings. Titles make an impression. They should be concise, to the point, and easy to say. Your title should be in larger text than the rest of your solution – perhaps at least 20pt if the body text is 10pt.

9780170440875



'Pesticide Planet' by G. Grullón/ Science, Science 16 August 2013: Vol. 341 no. 6147 pp. 730-731; DOI: 10.1126/science.341.6147730. Reprinted with permission from AAAS

FIGURE 1.38 'Pesticide Planet' has a large, visually interesting title.

Text styles

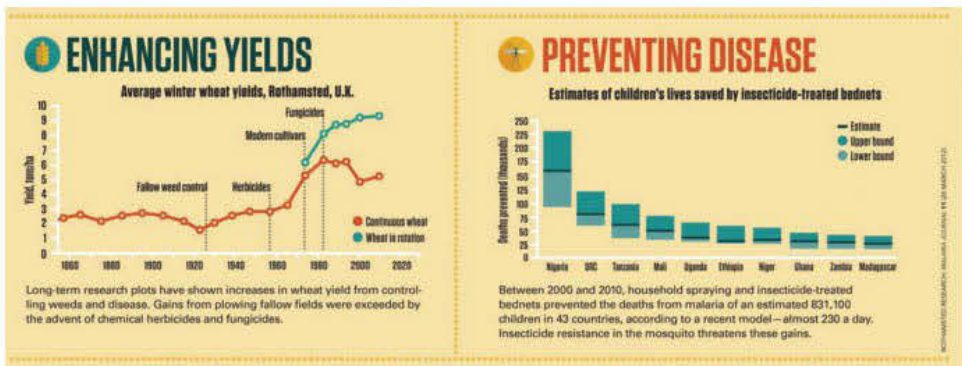
When we discuss text styles, we are talking about more than just fonts. A font is a typeface (such as Times New Roman, Arial or Calibri) plus its attributes (20pt, bold, red). You may already know a few standard, familiar typefaces, such as:

- Times New Roman, a serif typeface. **Serif** typefaces have tiny marks or 'tails' on the end of the horizontal and vertical strokes of each letter. Serifs are used in books for body text (such as in this book) and especially for long passages of text.
- Arial, a sans serif typeface. **Sans serif** typefaces do not have the serifs on the strokes of each letter. They work best for short paragraphs, large headings and online, but not for long passages of printed text.
- Courier New, a slab serif typeface (a sub-type of serif fonts), is often used in programming. Slab serifs are best used when the focus is on function and not appearance. Slab serif fonts ensure your characters are legible and unmistakable.

When you are choosing text styles for your graphic solution, keep things simple. Use a few well-chosen typefaces, perhaps three at most, and use bold, italic, colour and point size to set out heading levels and distinguish between different types of text.

The two sections in Figure 1.39 use the same three typefaces. The only difference is that the subheadings use different font colours and have icons to the left to distinguish between them.

There are other typeface styles, such as handwriting, script and decorative. You will know decorative fonts such as Impact because it is the typeface predominantly used in memes online. Comic Sans, a casual script or handwriting typeface, is also well-known.



'Pesticide Planet' by G. Grullón/Science, Science 16 August 2013: Vol. 341 no. 6147 pp. 730-731; DOI: 10.1126/science.341.6147730. Reprinted with permission from AAAS

FIGURE 1.39 Subtle changes make a difference in 'Pesticide Planet'.

Text styles will apply some contrast while promoting a streamlined, professional appearance. However, using many different typefaces in one space can be untidy and overwhelming.

Remember: Less is still more – bigger is not always better. Think about what needs to be emphasised the most and what needs to be highlighted. Not everything needs to be bold, italic and 40pt.

Colour and contrast

Colour should be used so that it makes the information clear, readable and attractive. The colours should emphasise important features, and a colour scheme should be used to ensure consistency. The following conventions for on-screen colour can be useful in determining colour schemes.

- The most easily readable colours for text are black writing on a white background.
- Avoid using red and green together because people who are colourblind have difficulty distinguishing between them.
- Blue and brown together can also be difficult to read.
- Shades are best used for backgrounds.
- Avoid using yellow or other light colours for text on a white background.
- Avoid using bright, neon or vivid colours, except where you wish to highlight an object or piece of information.
- Limit the number of different colours used in your graphic solution.

As discussed previously, contrast refers to the visual difference in colour or tone between objects in a graphic solution. Greater contrast will make objects appear to stand out more from one another. If there is not enough contrast between two objects, they may appear to blend into each other, making it difficult for the user to see each of them clearly. Contrast between the background of the graphic representation and text should make the information clearly visible and legible. The use of white space can enhance the contrast around objects within the graphic representation.

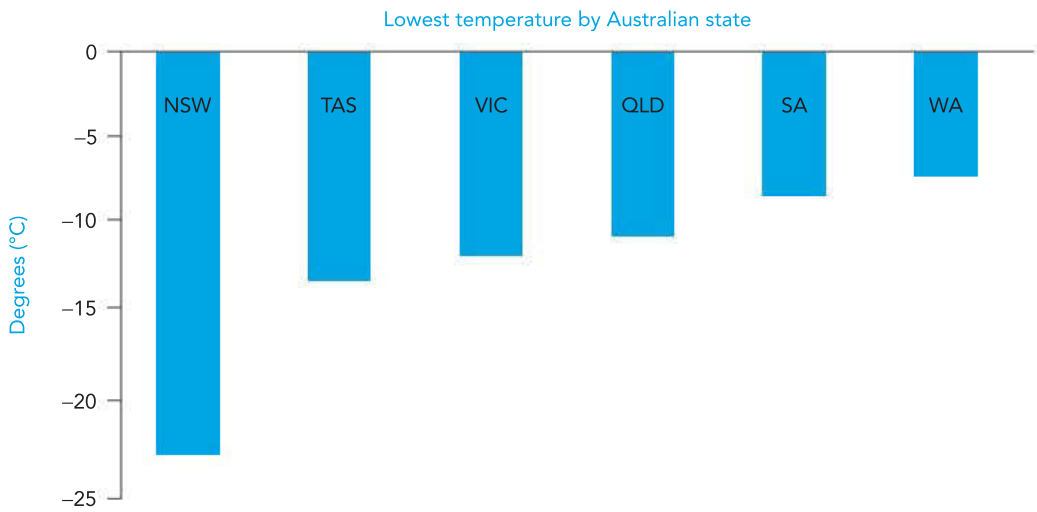


FIGURE 1.40 This graph shows use of clear contrast.

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

Shapes

Using shapes in your solution can help to create patterns, contrast, hierarchies and backgrounds. You can use shapes as containers for sections of text in your graphic solution, and as dividers.

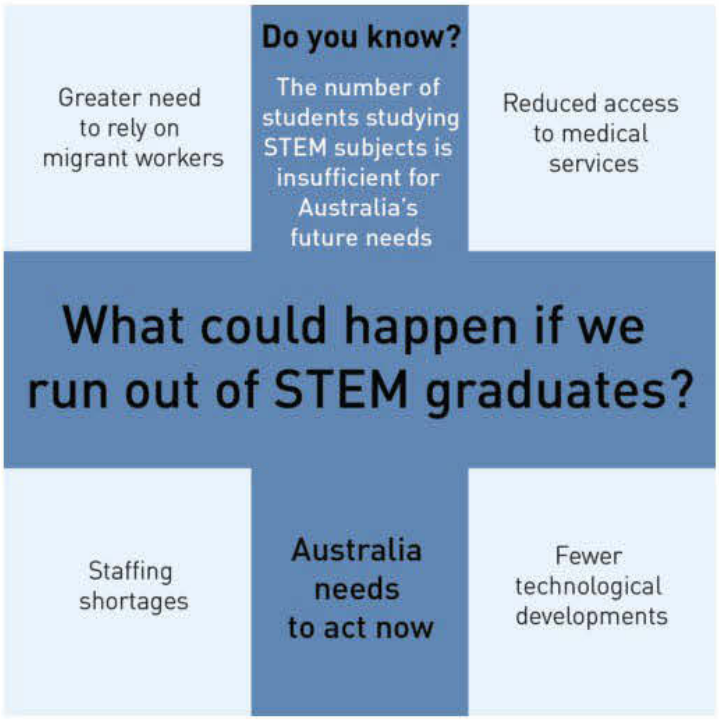


FIGURE 1.41 The cross shape here acts as a divider and container.

You do not have to stick to standard, two-dimensional geometric shapes such as squares, circles, rectangles and triangles. Other types of shapes, such as irregular, abstract and freeform shapes can evoke reactions in the user. You can also use shapes to develop logos, symbols and icons.

THINK ABOUT DATA ANALYTICS 1.6

1 Make a list of what you immediately associate with each of the shapes in Figure 1.42.
2 Compare your list with one of your classmates. How much do the lists overlap?

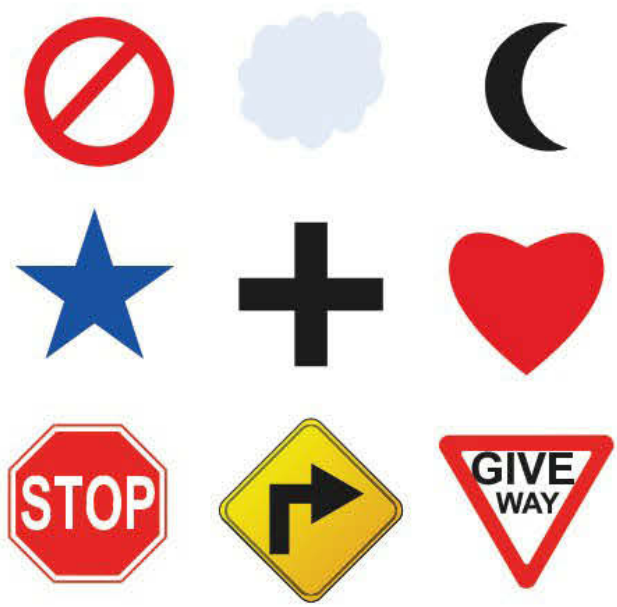


FIGURE 1.42 These familiar shapes may not be what you immediately think of when shapes come to mind, but they may still evoke strong reactions that could make them powerful in your graphic solution.

9780170440875

Lines and arrows

A line is a versatile visual element that uses only length and width. Lines can be any of the following.

- Bold or thick lines work well for emphasis and for representing a structure within a space. The thicker the line, the more it will draw the eye to the space, but the more crowded and boxed in it will look (so use a thick line carefully).
- Light and fine lines can suggest technical details but also retain a sense of minimalism.

Solid	
Dashed	
Dotted	
Broken	
Double	
Thick	
Thin	
Curved	
Freeform	

You can use lines in your graphic solution as borders or containers for sections of text or images.

The 'Pesticide Planet' infographic (Figure 1.43) uses dotted lines as dividers, but also as a form of repetition from the maps at the top. By using lightly coloured lines that are similar to the background colour, and dots rather than solid lines, the infographic allows the user to read the infographic in the correct order and tell sections apart, but without a sense of crowding that solid or darker coloured lines might have created.



Pesticide Planet' by G. Grullón/ Science, Science 16 August 2013: Vol. 341 no. 6147 pp. 730-731; DOI: 10.1126/science.341.6147.730. Reprinted with permission from AAAS

FIGURE 1.43 The dotted dividers add visual interest while serving the purpose of separating areas.

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

You can also use arrows as pointers in your graphic solutions. There are a variety of arrows and arrowheads to choose from, as seen in Figure 1.44.

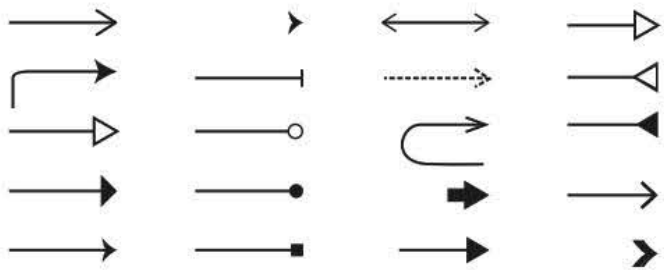


FIGURE 1.44 Popular arrows and arrowheads

Some arrows are sophisticated and elegant, while others are basic. The colour of your lines plus the colour of your arrow and choice of arrowhead can make a difference in the overall appearance of your diagrams.

Sources of data

If you are using data in your graphic solution, you need to identify the source in an appropriate way. If you are designing an infographic for your solution, you could include a list of all of your sources in the footer of the infographic (Figure 1.45).



FIGURE 1.45 Reference list

Alternatively, you could cite your source when it is used, similar to 'Pesticide Planet'. Note that 'Pesticide Planet' runs the source vertically up the side of the infographic (Figure 1.46, page 45).

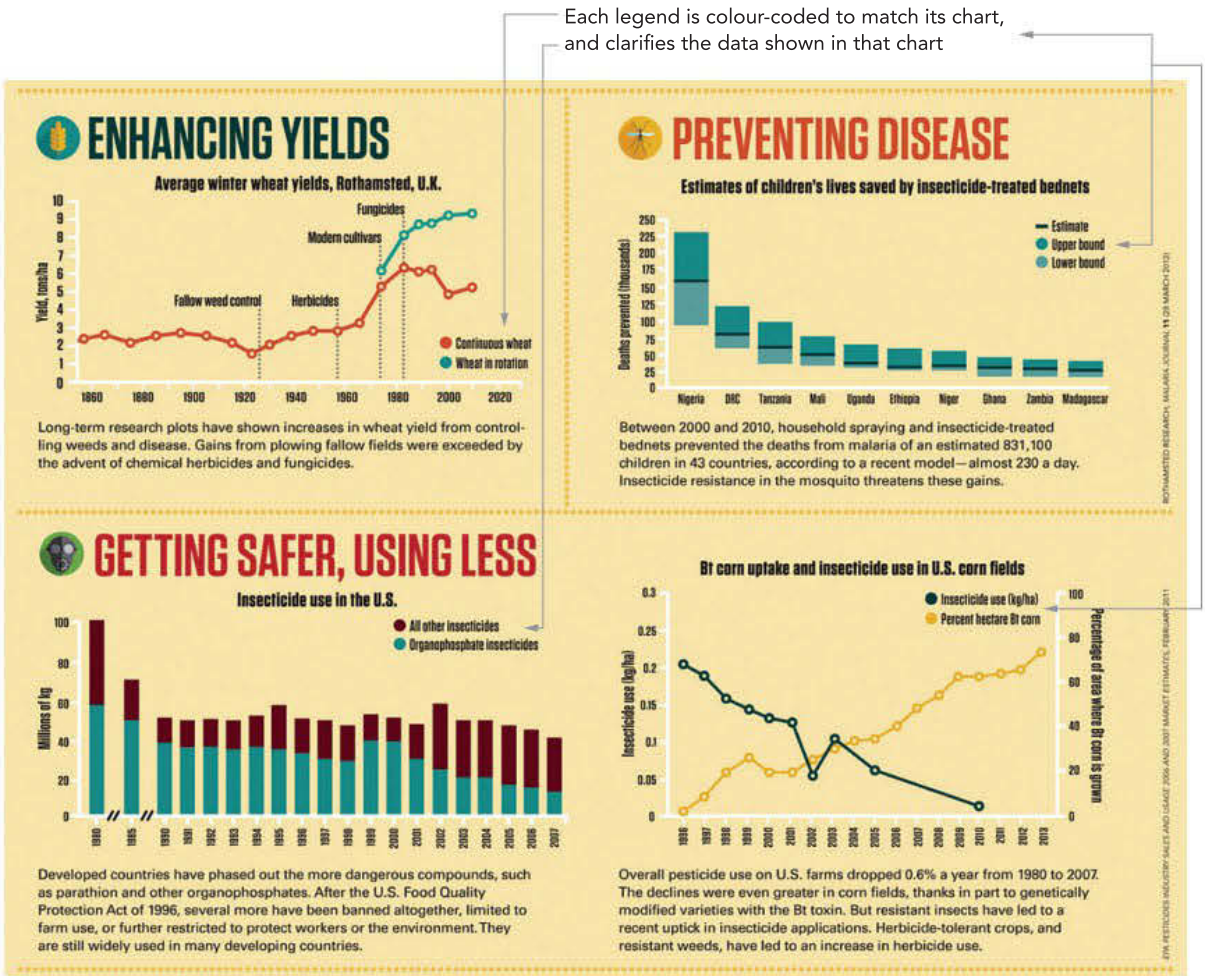
Citing information resources is a must when working with infographics because the data behind them is research-based. Citing sources also provides those who view the infographic the opportunity to further research the topic. Make sure you cite all of your sources correctly using the APA method described on page 14.

9780170440875

Legends

You should also make use of legends in your graphic solution when needed to identify the facts shown in charts or graphs clearly. In general, a legend or key explains the symbols used in a chart, diagram, map or table. In terms of your graphic solution, a legend will mostly be used as a patterned marker with blocks of colour that represent different groups of data in a chart.

'Pesticide Planet' includes multiple colour-coded legends – one for each 'module' that has a chart. The legends in 'Pesticide Planet' make it easy to understand what the data stands for and thus what each chart means, which is why legends are so useful. If you do not include a legend for a chart with a complex idea, the user may become confused about what is being shown on each axis and interpret your chart incorrectly. You can design legends to take up very little space, as shown in Figure 1.46.



'Pesticide Planet' by G. Grullón/Science, Science 16 August 2013: Vol. 341 no. 6147 pp. 730-731; DOI: 10.1126/science.341.6147.730. Reprinted with permission from AAAS

FIGURE 1.46 Legends in 'Pesticide Planet'

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

Design tools for data visualisation

Before you start to create any type of digital solution such as a data visualisation, you need to design it. The design process is part of the Data Analytics problem-solving methodology and begins after the *analysis* tasks have been completed. Their purpose is twofold: it helps you to focus and figure out what it is you want to build and how it will work and look, and it serves as a creation guide so that somebody else could create exactly what you have planned. Design tools cover all aspects of planning, from how the data is stored, manipulated and retrieved, how the navigation and user interface (UI) work to the solution's appearance.

This section will address only the tools used for data visualisations presented in a multimodal context.

Designing for data

Data dictionaries

The use of data dictionaries with databases will also be discussed in Chapter 2.

In this subject, **data dictionaries** are used to design the database tables that will hold your data. Their purpose is to establish the names of all tables and fields and to work out the properties for each field such as data type, field size, validation and labelling. All of these will be covered in detail in Chapter 2. Data dictionaries may also be used for designing the functionality of spreadsheets.

IPO charts

Input–process–output (IPO) charts are tools to help figure out what is needed to be done with data to create meaningful information. They are made up of three columns, labelled 'Input', 'Process' and 'Output'. To fill them out successfully, begin on the right-hand side – list the output you are expecting, as shown below in Table 1.4.

TABLE 1.4 Sample IPO chart with 'Output' column filled in

Input	Process	Output
		Total cost of tickets
		Image of logo without background

Once you have worked out what you want as a final result, figure out what raw data you need to be able to produce it, as shown in Table 1.5.

TABLE 1.5 Sample IPO chart with 'Input' column filled in, listing data required

Input	Process	Output
Cost of each type of ticket Number of tickets needed		Total cost of tickets
Original logo		Image of logo without background

Once that is completed, all that remains is to describe the process that will be used to convert the pieces of data (listed under input) into useful information (output). There are no set conventions here – you can describe it using prose (regular English text) or write a formula, both of which are modelled in Table 1.6.

9780170440875

TABLE 1.6 Sample completed IPO chart with 'Process' column finished

Input	Process	Output
Cost of each type of ticket Number of tickets needed	=NumAdultTix*AdultPrice + NumKidTix*KidsPrice	Total cost of tickets
Original logo	Import logo Remove background using photo editor Crop Save as .png	.png version of logo without background

Designing relationships

Data structure diagrams

Data structure diagrams (DSD) are used when designing databases to show how many tables are used and where they link together. They identify the table and field names, primary and foreign keys, and cardinality, all of which will be discussed in depth in Chapter 2.

Storyboards

Storyboards show how a user will navigate through an interactive solution, with a little overview of their appearance, like sketches of thumbnails. These are only used to show navigation and do not provide detail about the data or content of the pages.

Site maps

Site maps are hierarchical diagrams that show how web pages link together. They give an overall view of the site's pages, how they are structured and how they link together. They give a hierarchical view of the site, with the most important pages at the top, including the homepage. They include the page title and often provide the page's filename.

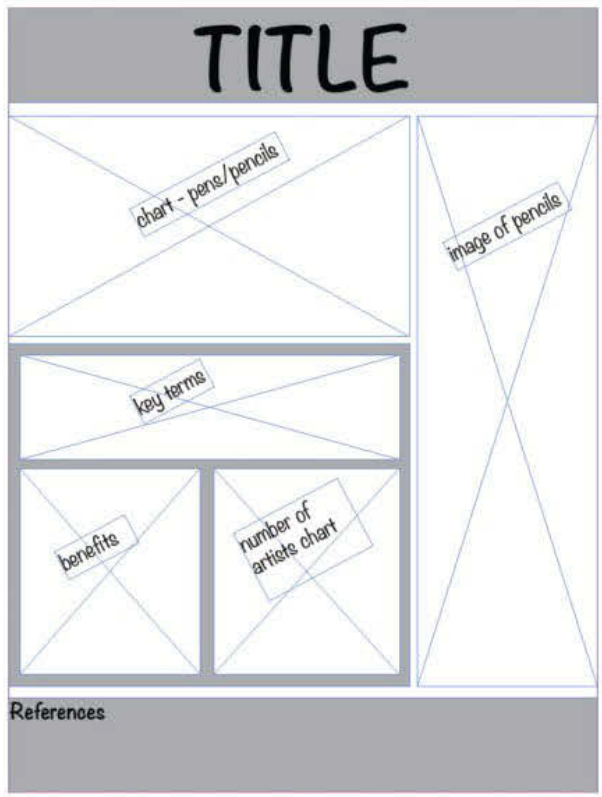


FIGURE 1.47 A wire-framed infographic about artists who use pencils

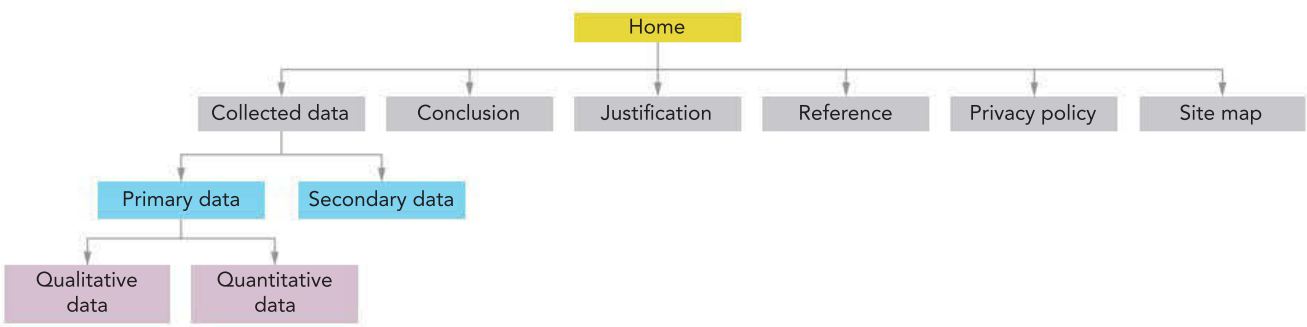


FIGURE 1.48 Site map

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

Designing appearance

It is important to give careful thought to how your solutions will look. This should make it easier to ensure that all required elements are included, that any restrictions (such as corporate, school or team colours or logos) are observed and that the client, where appropriate, can be consulted before the development begins. Designs also give the developers something to work from so that they can create the exact product that is desired.

Layout diagrams

Layout diagrams are rough sketches indicating where particular elements such as navigation, text boxes, images or charts should be placed on a page or web page. They often make use of placeholding text for text boxes (*lorem ipsum* being the most well-known) but often use simple boxes with crosses through them. This is known as wire framing. These diagrams are used to get a rough idea of what items will be placed where so that no time is wasted building an infographic (or other project) with elements in the wrong places.

Layout diagrams indicate features including:

- positioning of elements (buttons, scrollbars, charts)
- relative sizes of elements.

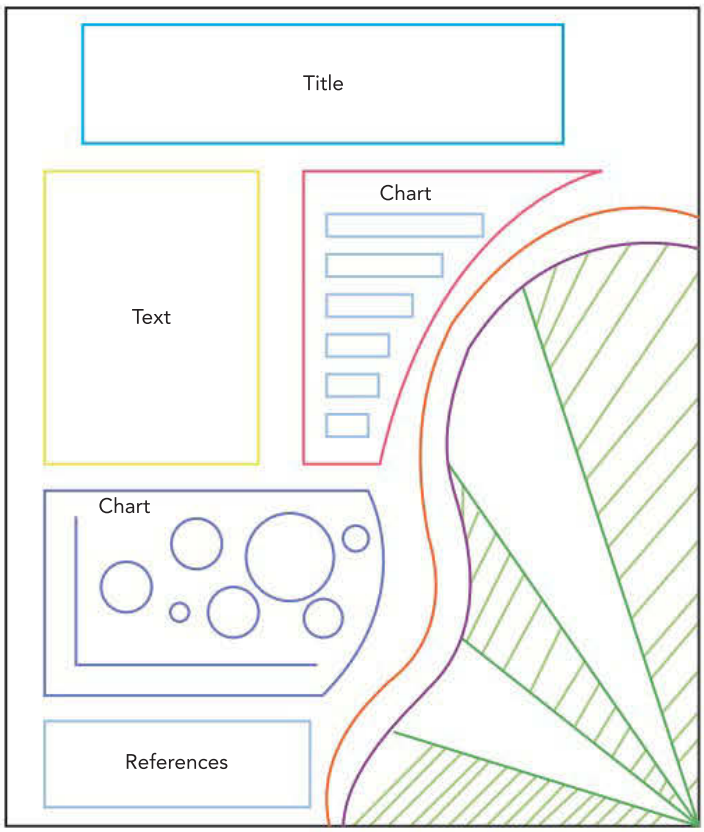


FIGURE 1.49 A mocked-up infographic

Mock-ups

Mock-ups are rough sketches of what a particular page, element, chart or interface might look like. They indicate features to show what something may look like so that potential clients can get an idea of what you are designing them and give feedback. Or they can be given to a developer so that they have a full understanding of what you want them to create. Mock-ups indicate features including:

- positioning of elements (buttons, scrollbars, charts)
- relative sizes of elements
- menu positioning and options
- alignment (vertical, horizontal or even diagonal)
- any appearance elements you wish to include (borders, images, colour choices).

Mock-ups do not include any extra text or annotations to explain them – if these are included, then they become annotated diagrams.

9780170440875

Annotated diagrams

Annotated diagrams share a lot of features already described in the previous two sections. Like layout diagrams, they show the positioning of objects without much visual detail. Like mock-ups, they do provide some more visual detail, but are usually closer to layout diagrams in this regard.

The big difference is that annotated diagrams are just that – annotated. They have notes on the side, or notes in the diagram itself explaining functionality, font and colour choices, where links go and any other details that would require further detail.

Annotated diagrams indicate features including:

- positioning of elements (buttons, scrollbars, charts)
- relative sizes of elements
- menu positioning and options
- alignment (vertical, horizontal or even diagonal)
- any appearance elements you wish to include (borders, images, colour choices)
- fine details about foregrounds, backgrounds, fonts, link appearances.

Remember: When annotating details about font choices, there are five decisions to be made:

- Font name
- Size
- Style (bold, italic)
- Colour
- Alignment

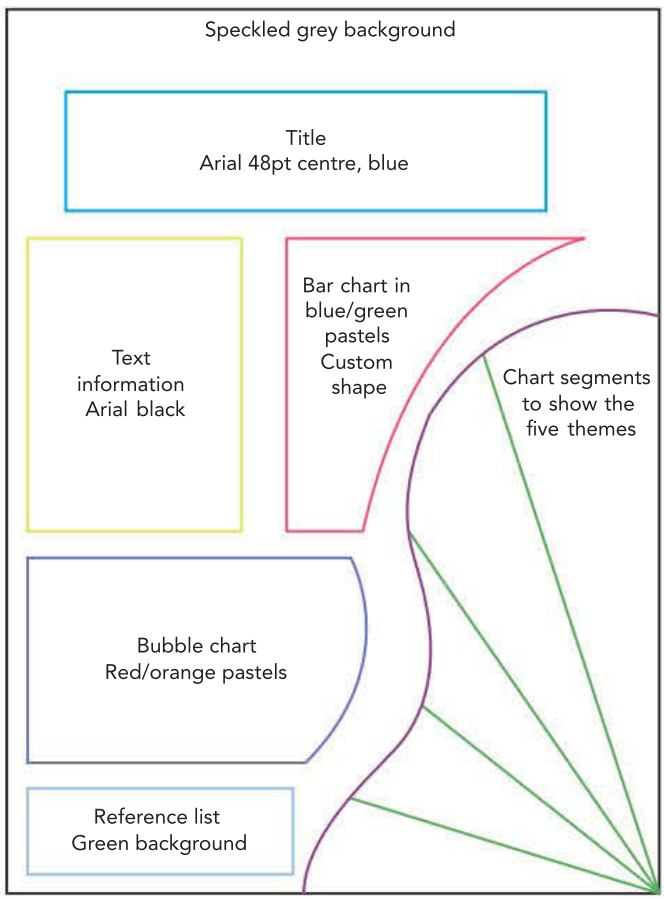


FIGURE 1.50 An annotated diagram for an infographic

Data collection, once validated and checked for integrity, is the beginning of the process when dealing with the creation of data visualisations. It is important to understand how these are combined with design tools to prepare for the manipulations that will be covered in Chapter 2. These processes will be used for both your Unit 3, Outcome 1 SAC as well as for your SAT.

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

1 CHAPTER SUMMARY

Essential terms

accuracy correctness, without errors (which contributes to data integrity)

acquisition gathering data

American Psychological Association (APA) referencing system used to acknowledge intellectual property

analysis first stage of the problem-solving methodology where problems are defined

authenticity data that is genuine, original and considered trustworthy (which contributes to data integrity)

bar graph visual tool showing totals as bars; the bars do not need to be related to each other

Boolean data a logical data type that can hold only two possible values, usually true/false, yes/no or 0/1

bubble chart visual tool showing numbers using circles to show their relative sizes as well as their relative positions on an *x* and *y* axis

character a single letter, number or symbol that is usually stored in 2 bytes of memory; a data type

chart a method of displaying data visually using symbols; sometimes known as a graph

citation a reference to the information's source

clarity something that is coherent and easily understood (which contributes to accuracy and data integrity)

clarity of message how well typical users interpret the information presented to them

column graph visual tool showing numbers as vertical bars

completeness the data has all the necessary parts (which contributes to accuracy and data integrity)

conditional formatting a way of changing the appearance of data to reflect values and make patterns easier to see

consistency treating data sets in the same way or ensuring that the appearance of items is formatted to look the same (which contributes to accuracy and data integrity)

constraint limitation on software solutions – can be economic, legal, technical, social and usability related

convention a way of doing things established by general consent or usage

copyright legal protection for intellectual property

correctness an element of data integrity to ensure the data is right

data raw, unprocessed facts and figures

database software tool for storing and manipulating large sets of data

data dictionary design tool for working out table and field names and properties

data integrity ensuring that all data is trustworthy, which is achieved through accuracy, authenticity, correctness, reasonableness, relevance and timeliness

data source where data can be found and downloaded (online)

data types particular forms that an item of data can take, including numeric, character and Boolean

data visualisation the presentation of data in a pictorial or graphical format

design second stage of the problem-solving methodology where designers ensure their work meets the requirements identified during analysis

design brief document outlining what is required in a solution

9780170440875

design principles expectations around what needs to be considered when designing functionality and appearance

effective a solution that performs as intended

efficient a solution that saves time, cost and effort

flow visualisation representing data that illustrates the flow pattern of an object or item

format how something is displayed (for example, using a table or chart); arranging the look or presentation of an object

Geographic Information System (GIS) a file format that contains geographical data

graph visual representation of data showing the relationships between several elements

hierarchy visualisation shows the relationships and structure between data items

histogram visual tool using bars to show distributions within a range

image data any stored images used to generate information

infographic visual representation of data, usually combining several charting and text tools

information useful knowledge created by manipulating data

information need when particular information is required, yet is not being currently supplied; causes of an information need may include a current problem, an identified need or an opportunity

information system software, hardware, people, processes and data that work together

input the process of entering data into a system

integer a number without a fractional or decimal component

intellectual property (IP) the ideas and works created by an individual or organisation

intended user target audience; that is, people expected to use the product

interview data acquisition technique that allows in-depth responses; direct questioning of a person

line graph visualisation tool that connects points to show continuous changes over time

line of best fit a straight line that represents the trend in a scatter plot

map-based visualisation geographical data displayed in a visual format

matrix visualisation visualisations used to show the composition of individual items in the sample size

mock-up a sketch of a solution's appearance

numeric data any data made up exclusively of numbers; includes integer and date

observation data acquisition technique that involves watching and gathering data in real time

pie chart visual representation showing an item broken up into its relative segments

plagiarism submitting or claiming someone else's work as your own

primary data new facts collected personally by a researcher to answer a specific question

problem-solving methodology (PSM) approach used to analyse, design, develop and evaluate projects or solutions

reasonableness data that is believable, but not necessarily accurate (which contributes to data integrity)

relevance data that produces useful information required by the information need; if the data is not useful, it is not relevant (which contributes to data integrity)

sample size the number of completed responses that you collect from interviews or surveys; the more responses completed, the more likely the data being collected is accurate

sans serif literally 'without serifs', which are the small lines at the ends of typefaces; 'sans' is the French word for 'without'

scatter graph visualisation to show a series of points on a graph

scope statement about what will and will not be included in a project

secondary data data collected by someone other than the researcher, which has often been processed

sensor equipment used to collect data electronically

9780170440875

sensor data data acquired or generated by machines monitoring particular conditions

serif the small line at the end of letters in typefaces such as Times New Roman

solution requirements what the client wants from the solution; can be broken down into functional and non-functional requirements

sound data any audio captures being used in a project

spreadsheet software tool for performing calculations and generating charts

stream graph visualisation that shows the relative levels of several variables over time

survey data acquisition technique involving a series of set questions, often completed online by large numbers of respondents

testing ensuring that something works as intended and that the outputs are accurate; that the outputs are correct with no errors, or that any errors can be dealt with without inconveniencing the user

text data type data that consists of a string of characters

timeliness data that has been collected in a reasonable timeframe (is not too old) and information that has been produced in time to be useful

time visualisation a visual that represents a data item or data set over a period of time

validation checking that data input is reasonable

weather data data produced about weather, climate or water

word cloud visualisation of data that is word-based (qualitative)

9780170440875

Important facts

1. **Data** is made up of raw, unprocessed facts and figures.
2. **Information** is derived from processing data into a form that humans can understand.
3. **Primary data** is first-hand; **secondary data** has been summarised previously in some way.
4. There are many government and non-government **public sources of data sets** available locally, nationally and internationally.
5. **Interviews** use open questions to gain detailed, personal opinions.
6. **Observation** gathers information about people acting naturally.
7. **Surveys** and questionnaires usually use closed questions to gather a lot of primary data.
8. **Data types** include number (integer, floating point), character (text, string), Boolean (true/false), image and sound. The data types of database fields must be chosen with care.
9. **Validation** should be used to check that data is reasonable as it is entered into a system. Checks include existence, data type and range.
10. **Data integrity** is critical to ensure that the data you are using is trustworthy. Data must be accurate, timely, authentic and relevant.
11. Use the **American Psychological Association (APA)** referencing system to acknowledge sources.
12. Proper referencing is needed to avoid **plagiarism**.
13. A **reference list** contains full details of every reference given in the body text.
14. The federal *Copyright Act 1968* protects the **rights of owners of intellectual property**, including electronic forms like software, websites and digital games, books, music, documents and movies.
15. **Conditional formatting** can make it easier to understand tables of data.
16. **Data visualisations** make numeric data easily understandable.
17. Different **chart types** serve different purposes. Know when to use histograms, bar charts, line graphs, pie charts, bubble charts, stream charts, and others to best convey your message.
18. Words cannot be converted to charts easily. They must either be converted to numbers first or be counted in some way to produce **word clouds**.
19. **Infographics** combine several visualisation tools to produce clear, visually interesting solutions to communicate a message.
20. **Designing** solutions before building them is vital to ensure the finished product solves the problem that is being addressed.
21. Appropriately selected **design tools** should be used. For data visualisations, these include data dictionaries and mock-ups.
22. **Formats** and **conventions** should be followed to ensure that users understand what they are looking at.

9780170440875

TEST YOUR KNOWLEDGE

Review quiz

Data

1 Identify the major difference between data and information.
2 Identify the main characteristics and strengths of primary data.

Acquiring data

3 Summarise the advantages of using interviews to collect data.
4 Explain one advantage that surveys and questionnaires have over observation and interviews.
5 List five reliable sources of large repositories of data. State why each can be trusted.

Referencing data sources

6 How does proper referencing avoid plagiarism?
7 Give two examples of using APA reference styles in body text and reference lists.

Data types

8 Why is it important to choose proper data types?
9 Which data type would you recommend be used for each of the following data sets? Justify your choice.
   a Pet breed
   b Pet date of birth
   c Registration status
   d Pet gender
   e Pet weight
   f Pet age
   g X-ray
   h Audio recording of heartbeats
10 Why is text the most commonly used data type in research?

Data integrity

11 Summarise the meaning of 'data integrity'.
12 How does lack of timeliness degrade the value of data?
13 How does incomplete data lead to faulty information?
14 Describe one way to discover if people are not telling the truth in surveys.
15 What does 'data authenticity' mean?
16 List three examples of how data may lose relevance.
17 What can be done to maintain the accuracy of data over time?

9780170440875

TEST YOUR KNOWLEDGE

Data visualisation

18 Why is data visualisation better than presenting tables of data?

19 List and describe six different types of charts, explaining when it would be appropriate to use them.

20 Define the term 'infographic'.

21 Describe the elements of an effective infographic.

22 List four different software tools that could be used to generate effective data visualisations.

Design

23 What is the purpose of a design brief?

24 Identify the design tool that would be used to describe data types and properties for a database table.

25 What is the difference between a mock-up and a layout diagram?

Formats and conventions

26 What is the difference between formats and conventions?

27 Identify the infographic in this chapter that you think has the most effective title and justify your choice.

28 Identify the infographic in this chapter that you think uses colours and contrasts least effectively. Justify your choice and suggest ways to improve the colour palette.

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875



APPLY YOUR KNOWLEDGE

1 You have been asked to design a project that will produce an infographic proving to non-believers that global warming is a reality.
   a Write a design brief (one or two paragraphs) that describes the problem you are facing, any constraints, and the scope of your problem.
   b Carefully select two sets of data that you can use.
      i Justify your choices and explain in detail why each set has data integrity. Make reference to all six elements of data integrity in your response.
      ii For each set of data, cite the web reference in the correct APA style.
   c Write 10 questions that you could use in a survey to gain an appreciation of how the people in your school community view global warming.
      i Explain the purpose of each question.
      ii Implement the survey with at least 10 people to gather data.
2 Draw a rough layout diagram of how your infographic would look. Make sure all important information is allocated space.
3 Create a mock-up of your infographic, showing the types of visualisations you think would best showcase the big ideas, community views (your classmates), referencing, and relevant statistics from reputable sources.

9780170440875

CHAPTER

2 Data manipulation and presentation

KEY KNOWLEDGE

After completing this chapter, you will be able to demonstrate knowledge of:

Approaches to problem solving

- naming conventions to support efficient use of databases, spreadsheets and data visualisations
- a methodology for creating a database structure: identifying entities, defining tables and fields to represent entities; defining relationships by identifying primary key fields and foreign key fields; defining data types and field sizes, normalisation to third normal form
- design tools for representing databases, spreadsheets and data visualisations, including data dictionaries, tables, charts, input forms, queries and reports
- design principles that influence the functionality and appearance of databases, spreadsheets and data visualisations
- functions and techniques to retrieve required information through querying data sets, including searching, sorting and filtering to identify relationships and patterns
- software functions, techniques and procedures to efficiently and effectively validate, manipulate and cleanse data including files, and applying formats and conventions
- types and purposes of data visualisations
- formats and conventions applied to data visualisations to improve their effectiveness for intended users, including clarity of message
- methods and techniques for testing databases, spreadsheets and data visualisations

Reproduced from the VCE Applied Computing Study Design (2020–2023) © VCAA; used with permission.

FOR THE STUDENT

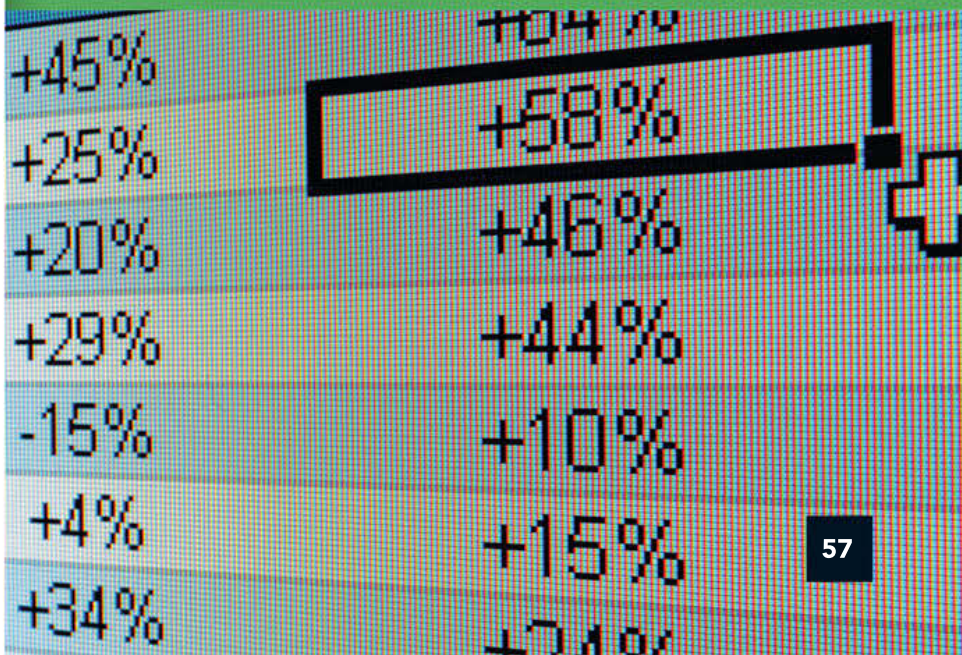
Acquiring large amounts of data is an important step in understanding the world around us and its problems. However, unless something is done with the data, it is largely useless. The data needs to be checked to see if it is reasonable, cleansed to remove any clearly unsuitable data, and then processed to extract exactly what is needed. This can then be further manipulated to create visual charts or other images to help the end-user quickly understand what the information is communicating.

This chapter will explain how you can cleanse data using a database and manipulate the data in a spreadsheet to make it ready for importing into a data visualisation tool. It will explain how databases are constructed and how spreadsheets are used for data preparation.

FOR THE TEACHER

This chapter introduces students to the knowledge and skills needed to use software tools including spreadsheets and databases to manipulate data into information, and to present that information in a readily understandable and attractive format.

The key knowledge and skills are based on Unit 3, Area of Study 1. Teachers will need to provide students with a design brief and designs using appropriate tools for students to build a database for cleansing data, a spreadsheet for further manipulation, and a data visualisation. If a data visualisation is effective, it enables readers to interpret information easily due to the clear and simple communication of its message.



iStock.com/slobo

9780170440875

Databases

THINK ABOUT DATA ANALYTICS 2.1

How many databases do you think you encounter each day? Which organisations store large amounts of data that need to be accessed in your daily routine?

A database is an organised collection of structured information, or data, typically stored electronically in a computer system. When correctly set up, they have minimal data redundancy, which makes the files relatively small and improves the integrity of the data contained therein.

Please note that Microsoft Access terminology and examples will be used in the following discussion, but all databases are structured in the same way, so if you are using another database you will still be able to understand the discussion that follows.

When to use databases

Databases can be used to organise large sets of data, especially where more complex **manipulations** are required. Spreadsheets are limited in the amount of data that can be stored. Although databases also have limitations, there are millions of records that are able to be stored.

Databases are also excellent tools for identifying problems in data, especially at the point of **input**, whether the entries are being typed in manually or imported. **Validation** rules will detect and quarantine any problems before they end up in the main database. These problems are kept separately so that the user can examine them closely and make informed decisions about how they should be handled.

A database with a single table is known as a **flat file database**. There is little that a flat file database can do that a spreadsheet cannot. As useful as a flat file database may be, it has limitations. The main limitation lies in the creation of redundant data. Often this takes the form of data that is repeated in each transaction, such as the mailing address of a customer or the details of a purchased product. To increase the efficiency and effectiveness of their data, organisations typically make use of **relational databases** instead of flat file databases. A software package written specifically to create these databases is called a relational database management system (or RDBMS).

CASE STUDY

Schools in Australia

Data acquisition

Carolyn and Julian were discussing the number of schools in New South Wales and Victoria. They wondered how many secondary schools there were. It was suggested that they check the ABS website for current numbers. They did this, but then they wondered whether that number had been changing over the years.

In order to find out, Carolyn and Julian need to locate raw data showing the total number of schools. This can be located on the ABS website.

Find a plain data file to download in .csv or .xls format. Avoid any data that has already been processed.

To acquire data sets, you will need to register on the ABS website. This is free. It is suggested you use your school contact details so that your purpose for using the data is clear.

9780170440875

To overcome the limitation of flat file databases, a RDBMS is made of multiple tables and also stores data **relationships** between tables. A relationship is a connection between the data. For a relationship to be established between two or more tables, they must have a common field. The primary key in a table usually acts as the field that joins the tables. When a primary key is used in another table, it is known as a foreign key.

A **one-to-one relationship** is used when a record in one table is connected to only one record in a second table. For example, an airline's passenger details table will contain records for many passengers, while a seat allocation table may hold records related to the seats on a particular flight. A one-to-one relationship exists between a passenger and their seat allocation. Each passenger has only one seat and each seat can be assigned to only one passenger.

Data in a one-to-one relationship can normally be accommodated on one table. Separate tables might be used if they have already been established for another purpose.

A **one-to-many relationship** indicates that one record in the first table can be connected to more than one record in a second table (Figure 2.1). For example, several workers in an office may share a single telephone extension. Each extension record is related to several employee records. The opposite of one-to-many is many-to-one.

How many is many? Consider a one-to-many relationship between individuals and cars. Each car is owned by only one person, but each individual may own one, five, twenty, or no cars. Thus 'many' can mean any number, including zero.

A **many-to-many relationship** is used when each record in the first table can be connected to a number of records in the second table. At the same time, each record in the second table may be related to many records in the first table. For example, a student detail table and a subject detail table may have a many-to-many relationship. Each student studies many subjects, and each subject is studied by many students. Another example of a many-to-many relationship is one in which an order from a customer for goods is related to a range of products. Each order can contain a number of products, while each product can appear on a number of different customer orders.

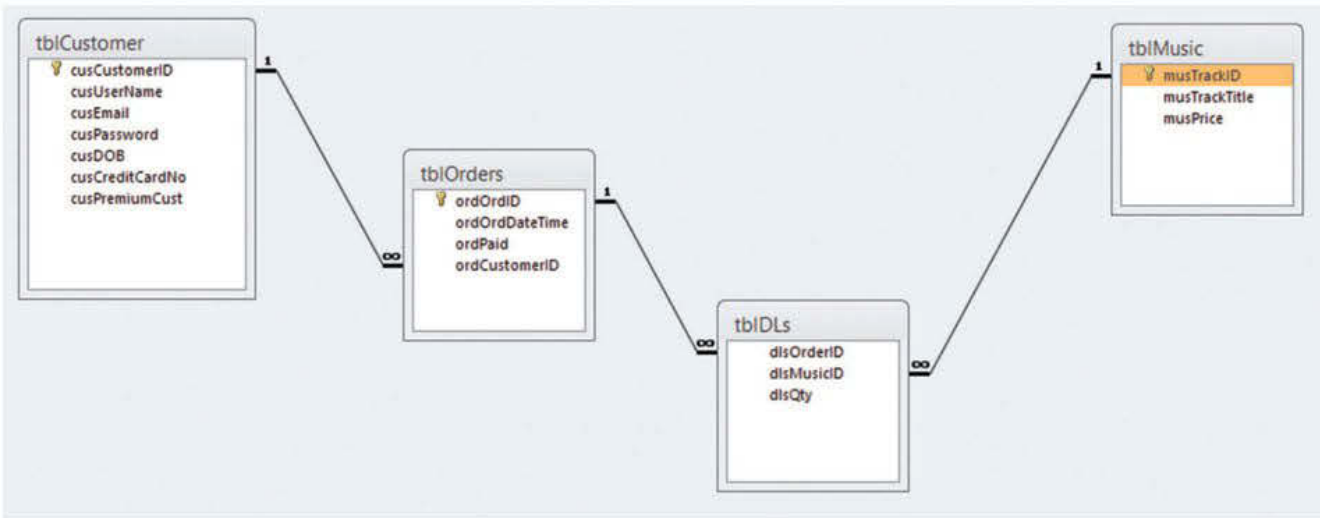


FIGURE 2.1 Tables linked by one-to-many relationships

Creating a database structure

Once you have identified that a relational database is required to solve an information problem, you will need to establish how to break up the data that will be supplied by the client into fields and tables. It is important to plan this carefully in order to maximise the efficiency of a relational database structure.

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

Entity relationship diagrams (ERDs) are often used to assist in establishing which data elements are required in a database (although they are not required for this course).

Determine the entities that exist in the problem. An **entity** is presented as a **table**, and is a single type of object about which data can be stored, such as people, places or things. Then the characteristics of each entity must be established. For example, these may include data elements like ID numbers, names, dates, addresses and prices. Each of the entity's characteristics would then be defined as a field. All the related fields of each entity would then be assigned to a table.

At this point, the arrangement of fields in tables may not be as efficient as possible. For this reason we would apply table normalisation (see below) to reduce the redundant fields and maintain data integrity.

The process of removing redundant data and arranging it into appropriate tables is called normalisation.

Within each table, the **primary key** would be identified to make the links with other tables where they will become **foreign keys**. The primary key must be a value that is unique for each record, such as a customer ID number. In terms of the relational aspect of the database, foreign keys (known as referential integrity in MS Access) are used to ensure that if you are entering data in one table, it already has a corresponding value in another table.

When you are happy with the table structures, establish the most appropriate data type and size (the number of characters to be stored in the field) for each field. To do this, examine the sample data supplied to you. This usually serves as a good indication of the general content for each field. At this point, you would be ready to create the RDBMS. The sections below follow this process and show you some of the design tools that will assist you in determining the structure of an RDBMS.

Table normalisation

Normalisation, in a nutshell, simply refers to relational databases that meet the following criteria.

- Each field only contains one piece of data.
- The fields in each table relate directly to that table's primary key.
- No calculated fields exist in the data tables.

Databases need to be thoroughly planned before they are built. It is recommended to break down the components 'on paper' into fields and tables. But the proposed table structure may not be as efficient as it potentially could be. So, at this point, table **normalisation** is employed. Normalisation is the process of ensuring that a database conforms to a set of normal forms. Its primary purpose is to remove redundancies that create threats to **data integrity** such as update anomalies. It also plays a role in making querying more efficient. There are six 'normal forms'. Each rule is applied successively from the **first normal form (1NF)** to the sixth normal form (6NF), but for the purpose of this study, you only need to apply the first three normal forms – first normal form (1NF), **second normal form (2NF)** and **third normal form (3NF)**.

Phone numbers, addresses and email addresses are partly redacted throughout this chapter for privacy reasons.

Normalisation works by providing a set of rules and a systematic procedure to check for various anomalies or deviations in data structure that would make the database less efficient, by ensuring that your fields are in the correct tables. It will not help if you have not chosen the correct fields in the first place.

These forms, in short, mandate that data must be broken down so that only one piece of data is in each field. The data should be split into logical tables and not have any fields dependent on another field. This will be discussed in greater depth on the next page.

A well-planned database will automatically comply with the normalisation rules. If the structure of the database does not fit into the normalisation rules, make adjustments to your design and try the rules again.

Figure 2.3 (page 61) shows a typical spreadsheet used to record orders. This could also be the single table in a flat file database. Spreadsheets are popular for holding raw data and for calculations because they are easy to construct and use, but their weaknesses become apparent when we start to ask more complicated questions of the data, such as 'How many of our Hawthorn customers bought products last week?', 'What is the value of orders placed by Widgets Inc. in the last financial year?' and 'What are our most popular items in each state?'

9780170440875

a

CustomerID	Name	Address	Suburb	Postcode	ContactPerson1	Telephone1	ContactPerson2	Telephone2
123	Widgets Inc	35 Colins Street	Melbourne	3000	Dawson, Bryn	0492 400 1	Nieves, Helen	0433 529 4
388	Poplets Pty Ltd	13a Mavis Road	Hawthorn	3122	Carter, Martin	0431 246 4	Clarke, Kareem	0489 440 5
50	Gears and Sprockets	Lot 2 Murray Road	Preston	3072	Haynes, Nina	0495 339 1	Fleming, Carolyn	0461 577 1

b

CustomerID	Name	Address	Suburb	Postcode	ContactPerson	Telephone
123	Widgets Inc	35 Colins Street	Melbourne	3000	Dawson, Bryn	0492 400 1
					Nieves, Helen	0433 529 4
388	Poplets Pty Ltd	13a Mavis Road	Hawthorn	3122	Carter, Martin	0431 246 4
					Clarke, Kareem	0489 440 5
50	Gears and Sprockets	Lot 2 Murray Road	Preston	3072	Haynes, Nina	0495 339 1
					Fleming, Carolyn	0461 577 1

FIGURE 2.2 Examples of unnormalised tables stored in a spreadsheet: a notice the repeating groups for contact person and telephone; b the same data is displayed and, even though it looks neater, it repeats groups vertically.

InvoiceNo.	Date	CustomerID	Name	Address	Suburb	Postcode	ItemID	ItemDescription	ItemQty	ItemPrice($)	ItemTotal($)	OrderTotalPrice($)
X239	10-Jan-16	123	Widgets Inc	35 Colins Street	Melbourne	3000	6673	Shovels - Long handled	5	25.95	129.75	
							7621	Hammer - 2lb	2	39.95	79.90	
							1523	M10 Screw - 12cm	80	1.25	100.00	
							492	Bucket - 9Lt	4	2.95	11.80	321.45
X240	11-Jan-16	388	Poplets Pty Ltd	13a Mavis Road	Hawthorn	3122	7621	Hammer - 2lb	3	39.95	119.85	
							6673	Shovels - Long handled	1	25.95	25.95	
							943	Rivets - bag of 200	2	9.99	19.98	165.78
X241	11-Jan-16	50	Gears and Sprockets	Lot 2 Murray Road	Preston	3072	1523	M10 Screw - 12cm	120	1.25	150.00	
							2884	Drill bits - assorted	8	2.59	20.72	170.72

FIGURE 2.3 A typical, unnormalised spreadsheet file used to store customer order details

To put this data into a form where we can easily and accurately get answers to these types of questions, we need to normalise in order to get the table and field structures needed to construct a proper RDBMS.

First normal form (1NF)

This rule states that there must be no repeating groups in the table. This means that no single row (record) contains more than one value in a field, nor will there be more than one column with the same kind of value. For example, in a table that lists products for online purchases, only the current price is recorded in the Price field, rather than the original price *and* the sale price. Where multiple telephone numbers exist for a customer, there are no Telephone 1, Telephone 2, Telephone 3 fields, only a single Telephone field (figures 2.2a and b).

Figure 2.3 is an unnormalised set of data. To get it into 1NF we need to first 'flatten' the table. While it looks like we are creating more duplication, this will all be removed by the time we reach 3NF. Figure 2.4 (page 62) shows the flattened Orders table.

We now need to identify a primary key (PK) for each record in the table. There is no single column (field) we can use, but if we combine the InvoiceNo. and the ItemID, then we can create a concatenated PK. For example, we might think of the concatenated PK for Row 2 as X2396673 and for Row 6 as X2407621. This table is now in 1NF.

A table's primary key is the smallest set of columns needed to uniquely identify a row in the table.

'Concatenated PK' here means to join together column values from a table. In this case from Figure 2.3, the concatenated primary key for Row 2 is derived from columns A and H.

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan | ☐ Collect complex data sets | ☐ Analysis | ☐ Folio of alternative designs | ☐ Infographic or dynamic data visualisations | ☐ Evaluation and assessment | ☐ Finalise report or visual plan

9780170440875

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	InvoiceNo.	Date	CustomerID	Name	Address	Suburb	Postcode	ItemID	ItemDescription	ItemQty	ItemPrice($)	ItemTotal($)	OrderTotal Price($)
2	X239	10-Jan-16	123	Widgets Inc	35 Colins Street	Melbourne	3000	6673	Shovels - Long handled	5	25.95	129.75	321.45
3	X239	10-Jan-16	123	Widgets Inc	35 Colins Street	Melbourne	3000	7621	Hammer - 2lb	2	39.95	79.90	321.45
4	X239	10-Jan-16	123	Widgets Inc	35 Colins Street	Melbourne	3000	1523	M10 Screw - 12cm	80	1.25	100.00	321.45
5	X239	10-Jan-16	123	Widgets Inc	35 Colins Street	Melbourne	3000	492	Bucket - 9Lt	4	2.95	11.80	321.45
6	X240	11-Jan-16	388	Poplets Pty Ltd	13a Mavis Road	Hawthorn	3122	7621	Hammer - 2lb	3	39.95	119.85	165.78
7	X240	11-Jan-16	388	Poplets Pty Ltd	13a Mavis Road	Hawthorn	3122	6673	Shovels - Long handled	1	25.95	25.95	165.78
8	X240	11-Jan-16	388	Poplets Pty Ltd	13a Mavis Road	Hawthorn	3122	943	Rivets - bag of 200	2	9.99	19.98	165.78
9	X241	11-Jan-16	50	Gears and Sprockets	Lot 2 Murray Road	Preston	3072	1523	M10 Screw - 12cm	120	1.25	150.00	170.72
10	X241	11-Jan-16	50	Gears and Sprockets	Lot 2 Murray Road	Preston	3072	2884	Drill bits - assorted	8	2.59	20.72	170.72

FIGURE 2.4 The Orders table has now been flattened. This table is now compliant with 1NF.

THINK ABOUT DATA ANALYTICS 2.2

2NF is simply about splitting data across multiple tables. What do you think would happen to a large database such as the Internet Movie Database (IMDb) if they did not use multiple tables?

Second normal form (2NF)

Put simply, this is where a table is in 1NF and any column (field) in that table that is not part of the actual PK must be wholly dependent on the concatenated PK. In other words, in situations where you have more than one PK field in a table, each non-key field must be fully dependent on both keys, not just dependent on one of the keys. To determine if tables comply with 2NF, look at each non-primary key field and determine if the field is wholly related to each of the PKs. If the field can exist independently of any of the PKs (that is, it is not wholly reliant on *all* the keys), then the table is *not* in 2NF. For example, in our Orders table from Figure 2.4, we have to see which fields depend on both the InvoiceNo. and ItemID columns. If we start with the Date field, we can say that it clearly relies on an order being made, so it is dependent on InvoiceNo. But an order date can exist without an ItemID, so immediately our table fails the 2NF check.

TABLE 2.1 What happens when the 2NF rule is applied to the other fields in our table from Figure 2.4. Of note are the fields that do not fit with either part of the concatenated PK, such as all the customer details, which will be dealt with in the 3NF process.

Column	InvoiceNo. (PK)	ItemID (PK)	Meets 2NF?	Comments
Date	✓	✗	✗	This only relates to InvoiceNo.
CustomerID	✗	✗	?	We will deal with this field in 3NF.
Name	✗	✗	?	We will deal with this field in 3NF.
Address	✗	✗	?	We will deal with this field in 3NF.
Suburb	✗	✗	?	We will deal with this field in 3NF.
Postcode	✗	✗	?	We will deal with this field in 3NF.
ItemDescription	✗	✓	✗	It never needs to be ordered.
ItemQty	✓	✓	✓	There is no quantity without an order and the quantity must be of an item.
ItemPrice($)	✗	✓	✗	Every item that exists must have a price, but the item doesn't have to be sold.
ItemTotal($)	✗	✗	✗	Both of these columns are derived from calculations, so they have no place in the table.
OrderTotalPrice($)	✗	✗	✗	

9780170440875

The next step is to take all the fields that depend in whole or in part on the second half of the concatenated PK (the ItemID field) and put them in their own new table. The other fields, including the ones we are unsure about, will remain in the existing table. Because this table no longer has a concatenated PK, it is now 2NF compliant. The 2NF checking process will be run again to see how the fields in the new Order Items table depend on the concatenated PK it has.

TABLE 2.2 Orders: What the table looks like at the end of this first part of the 2NF process

InvoiceNo. (PK)	Date	CustomerID	Name	Address	Suburb	Postcode
X239	10-Jan-16	123	Widgets Inc.	35 Colins Street	Melbourne	3000
X240	11-Jan-16	388	Poplets Pty Ltd	13a Mavis Road	Hawthorn	3122
X241	11-Jan-16	50	Gears and Sprockets	Lot 2 Murray Road	Preston	3072

TABLE 2.3 Order Items

InvoiceNo. (PK)	ItemID (PK)	ItemDescription	ItemQty	ItemPrice($)
X239	492	Bucket – 9 L	4	2.95
X239	1523	M10 Screw – 12 cm	80	1.25
X239	6673	Shovels – long-handled	5	25.95
X239	7621	Hammer – 2 lb	2	39.95
X240	943	Rivets – bag of 200	2	9.99
X240	6673	Shovels – long-handled	1	25.95
X240	7621	Hammer – 2 lb	3	39.95
X241	1523	M10 Screw – 12 cm	120	1.25
X241	2884	Drill bits – assorted	8	2.59

These are the two tables we end up with after the first part of the 2NF process. InvoiceNo. is brought across to the Order Items table so that we 'remember' the order to which each item belongs. The Orders table no longer has a concatenated PK because it is now based on the single InvoiceNo. field, but Order Items retains a concatenated PK.

In the second part of the 2NF process, we reapply the 2NF rules to the Order Items table. Again, we are looking for fields that depend on both parts of the concatenated PK.

We take out the fields that fail 2NF and create a separate table for them. This new table will be called 'Items'. The ItemQty field stays in the Order Items table (it moved in the previous step). It is now 2NF compliant because there is now a many-to-one relationship between the Order Items and Items tables (in relation to the Invoice table, it has a one-to-many relationship so it had to move).

Table 2.5, Table 2.6 and Table 2.7 are the final versions of the tables for 2NF.

9780170440875

TABLE 2.4 The results of reapplying the 2NF rules to the Order Items table

Column	InvoiceNo. (PK)	ItemID (PK)	Meets 2NF?	Comments
ItemDescription	✗	✓	✗	It never needs to be ordered.
ItemQty	✓	✓	✓	There is no quantity without an order and it has to be a quantity of an item.
ItemPrice($)	✗	✓	✗	Every item that exists must have a price, but the item doesn't have to be sold.

TABLE 2.5 How the tables look at the completion of 2NF (Orders table, no change)

InvoiceNo. (PK)	Date	CustomerID	Name	Address	Suburb	Postcode
X239	10-Jan-16	123	Widgets Inc.	35 Colins Street	Melbourne	3000
X240	11-Jan-16	388	Poplets Pty Ltd	13a Mavis Road	Hawthorn	3122
X241	11-Jan-16	50	Gears and Sprockets	Lot 2 Murray Road	Preston	3072

TABLE 2.6 How the tables look at the completion of 2NF (Items table)

ItemID (PK)	ItemDescription	ItemPrice($)
492	Bucket – 9 L	2.95
943	Rivets – bag of 200	9.99
1523	M10 Screw – 12 cm	1.25
2884	Drill bits – assorted	2.59
6673	Shovels – long-handled	25.95
7621	Hammer – 2 lb	39.95

TABLE 2.7 How the tables look at the completion of 2NF (Order Items table)

InvoiceNo. (PK)	ItemID (PK)	ItemQty
X239	492	4
X239	1523	80
X239	6673	5
X239	7621	2
X240	943	2
X240	6673	1
X240	7621	3
X241	1523	120
X241	2884	8

9780170440875

Third normal form (3NF)

This is where a table is in 2NF and any column that is not part of the primary key is dependent only on the primary key and no other column. Therefore, to be 3NF compliant, every field in a table must relate directly to the PK. In our 'Ordering' example, having established that the fields in the Items and Order Items tables are fully dependent on their own respective PKs, we come to the fields related to the customer details in the Orders table. We know that an order date cannot exist without an InvoiceNo., so that will be left in its existing table. But customers and their own details can exist separately to an InvoiceNo. A customer can exist, but does not have to make an order; therefore, the customer details fail 3NF because they are not dependent on the PK (InvoiceNo.) in their existing table. They can, however, be put in their own table because all those fields relate directly to the CustomerID field as their PK. If we split the Orders table to create a Customers table, the resulting tables will look like tables 2.8 and 2.9.

THINK ABOUT DATA ANALYTICS 2.3

3NF is essentially making sure that there is no field that depends on another field, such as a calculated field that works out somebody's *age* from their *Date of Birth* field. Why do you think it is a good idea to do this? **Hint:** What happens to the data a year after it has been entered?

TABLE 2.8 Orders

InvoiceNo. (PK)	Date
X239	10-Jan-16
X240	11-Jan-16
X241	11-Jan-16

TABLE 2.9 Customers

CustomerID (PK)	Name	Address	Suburb	Postcode
123	Widgets Inc.	35 Colins Street	Melbourne	3000
388	Poplets Pty Ltd	13a Mavis Road	Hawthorn	3122
50	Gears and Sprockets	Lot 2 Murray Road	Preston	3072

In the Orders table and the newly created Customers table after the first stage of 3NF, our tables are now taking shape, but there is a problem. By splitting them, we have also removed their relationship with one another because an order needs a customer connected to it, even though customers do not need to have any orders. To restore this relationship, we will place a foreign key (FK) in the Orders table. The purpose of the FK is to point to a PK in another table. In this case, the field concerned is CustomerID. You may also notice that the InvoiceNo. and ItemID fields are now also listed as FKs, but the database will concatenate them to form a PK for the Order Items field. This is a small refinement, but explains why we have not needed to create a specific PK for the Order Items table.

Table 2.10, Table 2.11, Table 2.12 and Table 2.13 (page 66) are the final version of our tables. They are now fully compliant to 3NF.

9780170440875

TABLE 2.10 Orders

InvoiceNo. (PK)	Date	CustomerID (FK)
X239	10-Jan-16	123
X240	11-Jan-16	388
X241	11-Jan-16	50

TABLE 2.11 Customers

CustomerID (PK)	Name	Address	Suburb	Postcode
50	Gears and Sprockets	Lot 2 Murray Road	Preston	3072
123	Widgets Inc.	35 Colins Street	Melbourne	3000
388	Poplets Pty Ltd	13a Mavis Road	Hawthorn	3122

TABLE 2.12 Items

ItemID (PK)	ItemDescription	ItemPrice($)
492	Bucket – 9 L	2.95
943	Rivets – bag of 200	9.99
1523	M10 Screw – 12 cm	1.25
2884	Drill bits – assorted	2.59
6673	Shovels – long-handled	25.95
7621	Hammer – 2 lb	39.95

TABLE 2.13 Order Items (these become a concatenated PK)

InvoiceNo. (FK)	ItemID (FK)	ItemQty
X239	492	4
X239	1523	80
X239	6673	5
X239	7621	2
X240	943	2
X240	6673	1
X240	7621	3
X241	1523	120
X241	2884	8

9780170440875

Database components

Databases are different to other data management software applications in that they are built up using several components, all of which serve specific purposes.

Naming conventions

Databases are traditionally created using **naming conventions**. As discussed in Chapter 1, conventions are the expectations around how items should appear or be formatted. Naming conventions are another type of expectation; in this case, around how components of a database are named.

Descriptive names

The most important thing is to name each element in a way that accurately describes its content or function. Short names are better. For example, a field for containing people's surnames should be called Surname.

TABLE 2.14 Default field names make it difficult for the user to instantly see what each one should contain. Which of Field2 and Field3 is the given name or surname? What is Field4? If Field4 was named *Weight*, then it would be much clearer.

Field1	Field2	Field3	Field4
Mr	Tom	James	82

Title	GivenName	Surname	Weight(kg)
Mr	Tom	James	82

No spaces

It is important when naming database elements that no spaces be included. This is important especially if a programmer later on needs to write code to interact with the database. While databases will accept spaces in names, it is not conventional to include them.

Capitalisation

If a name is to be made up of two or more words, it can be difficult to read them without spaces. To assist the user, capitalising each new word can make them much easier to read. This is referred to as *camel casing*. For example, a field for people's given names might be *GivenName*.

Use of prefixes

Many people choose to use what is referred to as Hungarian notation to more fully describe their database elements. It involves using a three-letter lowercase prefix to describe the object or data type. This approach is used widely in programming. Common prefixes include tbl (table), frm (form), qry (query), rpt (report), txt (text), int (integer), btn (button) and lbl (label).

9780170440875

CASE STUDY

Schools in Australia

Designing tests

As you design each element of your solution, you need to think about how you can test that each item works as expected. This testing must be documented. The best way to formally document test is to set up a **testing table**. At the design stage, it is important to determine what will be tested, how it will be tested (including input data in testing calculations) and what the expected result will be if the element works as expected. This part is crucial: if you don't know what it should do, then how can you know if something needs to be fixed?

Set up the testing table with columns as shown in Table 2.15 below. Work out which tests you want to run. Good choices include testing layouts, calculations, interactivity such as buttons or links, and queries that display results correctly. The last two columns are left blank during the design phase because you have not yet built the solution and therefore cannot test it. As you create each element, you can test them to see if they work.

TABLE 2.15 Sample testing table showing one element from each of the database, spreadsheet and visualisation parts of the case study. You will need to test multiple elements for your SAC and SAT. The completed version of this table can be viewed later in this chapter (Table 2.27).

Item tested	How it was tested (data)	Expected result	Actual result	How it was fixed
Database query: Only records with Vic, NSW, secondary schools and 2016–2017 are displayed	NSW, 2016, secondary Vic, 2017, secondary NT, 2017, secondary 2015 Primary	Accepted Accepted Rejected Rejected Rejected		
Spreadsheet: Formula in cell F8 works	Manually check and compare	10		
Data visualisation	Check chart for clarity	Easy to interpret		

Tables (Entities)

Tables, also known as entities, in databases serve one main function – to store the data. They are created with parameters outlining what each field can and cannot contain (validation) and they can be connected or linked to each other.

As already discussed, relational database management systems (RDBMS) are simply databases where there is more than one table linked together. Tables also have several components to them, and it is important to understand what each of them is and what their purpose is.

Fields

A **field** is the smallest part of a table. The equivalent in a spreadsheet is a cell. Each only contains one piece of data. Fields need to be defined in databases, and it is best to define them before data is imported. Decisions must be made for each field, choosing properties that include the following.

- Field name – short, descriptive and no spaces (this comes in useful if you need to do any programming with your database later on)

9780170440875

- Data type – as discussed in Chapter 1 (pages 16–18)
- Field length (if appropriate) – choose a value a bit longer than your longest entry; this reduces database size since any part of the character allowance that is left blank is stored as a space
- Format (if appropriate) – generally applies to how number fields are displayed
- Validation
- Required field – whether or not the user must enter a value (use sparingly)
- Input mask – a set format for data entry (for example, DD-MM-YYYY as opposed to MM-DD-YY)
- Primary key

Key fields

Primary keys are essential in relational databases. A primary key is a field that uniquely identifies a record. It is used to connect every field that has been entered together, so it is essential that no two records have the same primary key. Usually, an ID field is added to specifically address this purpose, although sometimes there could be a field that might be suitable. If there is any possibility of there being two entries with the same primary key, then a different field must be selected or created. For example, surnames would not be suitable because it is quite feasible that more than one record may have the same surname; similarly, telephone numbers and even email addresses are often shared by more than one person. Primary keys are required fields – they cannot be left blank because they are the essential identifier.

Foreign keys are fields in tables that are used to link to primary keys in other tables. They must exactly match the properties of the primary key to which it will link. The name of the field must be absolutely identical in order for links to be formed. They must also be the same data type.

It is crucial when planning databases that thought is given to how links will be created and that the fields that must match do so.

Validation rules

Validation can and should be added to database table fields where appropriate. While it is tempting to add validation to every field, it is important to remember the end-user when building tables. It is possible that they may wish to exit the form they are completing, but if there are too many restrictions on the data, the software solution may not allow them to exit until all validation conditions have been met. To avoid unnecessary frustration for the user, only use validation on fields where it is absolutely needed. See Table 2.16 (page 70) for examples of validation options.

Records

A **record** is a collection of fields that are all related to each other. They all connect to a single primary key, which is used to identify the record. If looking at a database table, each record is contained in one row, filling in one entry for each field. For example, in a database table of people, one record might consist of one each of a person's first name, surname, address, postcode, state, phone number and date of birth.

9780170440875

TABLE 2.16 Database validation options and examples

Validation type	Example(s)	Comment
Existence check	Required field	If you activate a required field, users will not be permitted to click out of the field until something has been entered. Use this sparingly since it can be frustrating for the user and they may enter false information simply so that they can click out.
Type check	Choosing a data type	This is in-built into most database software. For instance, selecting Boolean would limit users to only selecting *yes* or *no* responses (or similar).
Range check	Between 10 and 50 >10	This limits the user to only enter values that are within an acceptable range. Be careful of the boundary values – decide in advance if the boundaries should be included or excluded, and make sure you fully test this.
Input mask	0000	This only allows four-digit numbers to be entered.
	LLL	This only allows three letters to be entered.

THINK ABOUT DATA ANALYTICS 2.4

Why do you think it is important to design software solutions before building them? What benefit would there be to an organisation in doing this?

Designing tables

When planning a database, the most important thing is to properly plan the tables using a data dictionary. It doesn't matter how you plan to enter the data, the validation that databases can carry out is powerful if done correctly. For manual entry, each error that is picked up will receive an alert. If importing large amounts of data, as you will be doing, it will isolate the entries that do not meet the conditions you set. Unlike spreadsheets, there are no decisions to be made concerning appearance, since database tables are not intended to be seen by the end-user (this will be covered later, under 'Forms' on page 77).

Designing database tables takes time, but it is essential. You must decide on several things, including field names, data types and enough details about each so that somebody else could build the table exactly as you intended it. Choosing a field size is critically important since the default size is often 255 characters. Databases will always allocate the exact required size to every record to make retrieval of records much faster, so if your field only contains two letters, such as *Mr*, it will store the *M*, the *r* and 253 blank spaces. When multiplied by millions of records, this makes the database files huge. It is good practice to limit all text field sizes to a reasonable number (think long, double-barrelled surnames) and then add 5–10 characters just in case.

Table 2.17 demonstrates how a data dictionary should look. The table gives a sample data dictionary for a coin collection table called *tblCoins*. Not every cell in the table needs to be filled in; if it is not relevant, leave it blank. Note that there are no calculated fields (such as age of coin).

9780170440875

TABLE 2.17 *tblCoins*

Field name	Data type	Size	Caption	Format	Validation	Validation text	Other
coiCoinID	Autonumber				Required		Primary key
coiCountry	Text	30	Country of origin		Required		
coiDiameter	Number	30	Diameter in mm	Integer	Between 10 and 50	Please enter a value between 10 and 50	
coiMetal	Text	20					Lookup from linked table
coiYear	Number (int)				Input mask ----	Please enter a four-digit year	
coiDateAcq	Date		Date of acquisition	Short date	<=Now()	Cannot be a date in the future	
coiValue	Number						
coiCurrency	Text	15					
coiShape	Text	10		Lookup value from list			
coiValid	Boolean		Is the coin current and valid?	Checkbox			

Schools in Australia

CASE STUDY

Part A: Designing tables

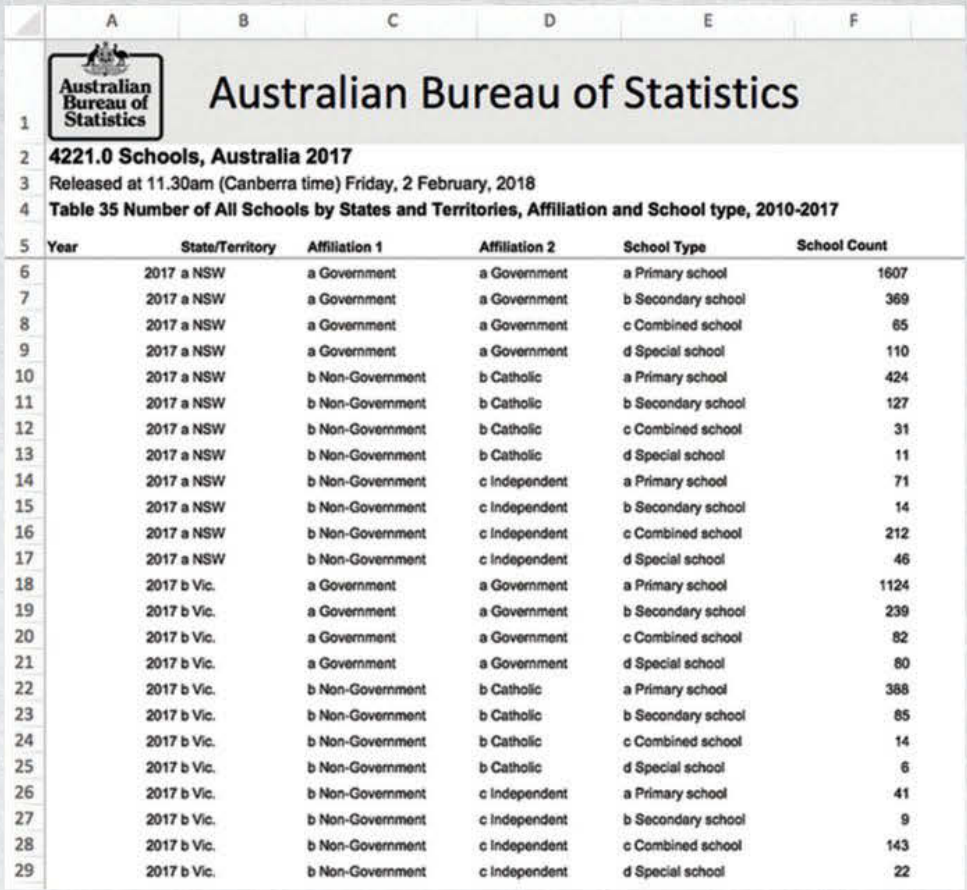
Set up your database table ready for your data. Sometimes it may be better to open the file in a spreadsheet first to move some fields around.

The data may appear in a spreadsheet such as Figure 2.5, page 72.

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875



Australian Bureau of Statistics

4221.0 Schools, Australia 2017

Released at 11.30am (Canberra time) Friday, 2 February, 2018

Table 35 Number of All Schools by States and Territories, Affiliation and School type, 2010-2017

Year	State/Territory	Affiliation 1	Affiliation 2	School Type	School Count
2017	a NSW	a Government	a Government	a Primary school	1607
2017	a NSW	a Government	a Government	b Secondary school	369
2017	a NSW	a Government	a Government	c Combined school	65
2017	a NSW	a Government	a Government	d Special school	110
2017	a NSW	b Non-Government	b Catholic	a Primary school	424
2017	a NSW	b Non-Government	b Catholic	b Secondary school	127
2017	a NSW	b Non-Government	b Catholic	c Combined school	31
2017	a NSW	b Non-Government	b Catholic	d Special school	11
2017	a NSW	b Non-Government	c Independent	a Primary school	71
2017	a NSW	b Non-Government	c Independent	b Secondary school	14
2017	a NSW	b Non-Government	c Independent	c Combined school	212
2017	a NSW	b Non-Government	c Independent	d Special school	46
2017	b Vic.	a Government	a Government	a Primary school	1124
2017	b Vic.	a Government	a Government	b Secondary school	239
2017	b Vic.	a Government	a Government	c Combined school	82
2017	b Vic.	a Government	a Government	d Special school	80
2017	b Vic.	b Non-Government	b Catholic	a Primary school	388
2017	b Vic.	b Non-Government	b Catholic	b Secondary school	85
2017	b Vic.	b Non-Government	b Catholic	c Combined school	14
2017	b Vic.	b Non-Government	b Catholic	d Special school	6
2017	b Vic.	b Non-Government	c Independent	a Primary school	41
2017	b Vic.	b Non-Government	c Independent	b Secondary school	9
2017	b Vic.	b Non-Government	c Independent	c Combined school	143
2017	b Vic.	b Non-Government	c Independent	d Special school	22

4221.0 Schools, Australia 2017. Released at 11.30am (Canberra time) Friday, 2 February, 2018. Table 35 Number of All Schools by States and Territories, Affiliation and School type, 2010 – 2017, http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/4221.02017?OpenDocument

FIGURE 2.5 School numbers data

In this case, if the data were being manually entered, a relational database would be set up using multiple tables (such as tables 2.18a–e).

TABLE 2.18a *tblYear*

Field name	Data type	Size	Caption	Format	Validation	Validation text	Other
yeaYear	Text	4			Required Unique Input mask: 0000		Primary key

TABLE 2.18b *tblState*

Field name	Data type	Size	Caption	Format	Validation	Validation text	Other
staState	Text	4	E		Required Unique		Primary key

TABLE 2.18c *tblSchLevel*

Field name	Data type	Size	Caption	Format	Validation	Validation text	Other
levSchLevel	Text	9	Primary/ secondary		Required Unique		Primary key

TABLE 2.18d *tblSchType*

Field name	Data type	Size	Caption	Format	Validation	Validation text	Other
typSchType	Text	12	School type		Required Unique		Primary key

TABLE 2.18e *tblSchCounts*

Field name	Data type	Size	Caption	Format	Validation	Validation text	Other
couCountID	Autonumber				Required Unique		Primary key
couYear	Text	4			Required Unique Input mask: 0000		Foreign key – links to *tblYear*
couState	Text	4	E		Required Unique		Foreign key – links to *tblState*
couSchLevel	Text	9	Primary/ secondary		Required Unique		Foreign key – links to *tblSchLevel*
couSchType	Text	12	School type		Required Unique		Foreign key – links to *tblSchType*
couCount	Number		Number of schools	Integer	Between 0 and 1000	Please check the number again	

However, if you look at the way the data is actually structured, it is in a flat file table, so the data dictionary for a direct import (bringing in the data in bulk, rather than typing each record individually) would look like Table 2.19.

TABLE 2.19 Data dictionary

Field name	Data type	Size	Caption	Format	Validation	Validation text	Other
couCountID	Autonumber				Required Unique		Primary key
couYear	Text	4			Input mask 0000	Please enter a four-digit year	
couState	Text	15					
couType	Text	20					
couType-Specific	Text	20					
couLevel	Text	15	Primary, secondary or combined				
couCount	Number		Number of schools				

9780170440875

Create the database table, making sure all field types and properties match the data dictionary in Table 2.19.

tblCount

Field Name	Data Type
ID	AutoNumber
CalYear	Short Text
State	Short Text
Type	Short Text
Type-Specific	Short Text
Level	Short Text
Count	Number

FIGURE 2.6 Completed table (in Microsoft Access design view) (note that ID is set as the primary key)

Part B: Importing data

Change the headings in the spreadsheet file to match the ones in your database table. Import the data file into your chosen database software.

Some databases may offer to normalise the data for you by creating tables. You need to be aware of how your data is being treated to make sure its integrity is not being compromised.

Note: Some data sets may be protected or locked when you download them. If a password is required to unlock them, you may not be able to make changes to the sheet or data. Unlocking in Excel can easily be done using *File > Protect Workbook* if you have the correct password.

Designing relationships between tables

Relational databases require tables to be linked together. There are a few different ways to document designs for showing these links, but the most common is an entity relationship diagram. Again, there are a few different types of these diagrams, but as long as the intent is clear, it doesn't matter which is used.

Remember, the idea of linking tables is to remove data **redundancy**. When linking tables, it is important that the field used to link the two tables is identical in both tables, and the primary key is in at least one of them. The field at the end of the link that is not a primary key is referred to as a foreign key.

Queries

Queries are where databases manipulate data into useful information. They access data stored in tables and perform any number of tasks on it. They can perform calculations, hide or show fields, **search**, **sort** or filter fields and combine multiple fields together. Every time a query is performed, it displays only the records that meet the set criteria. This does not delete or remove records, it just hides the ones that are not relevant.

Searching

Searching databases locates every record that contains the exact string (sequence of letters) or numeric parameters (such as particular values or ranges). This is helpful when the user might require different information each time they use it.

9780170440875

Sorting

Sorting records is useful to help users quickly locate particular records in a list. Sorts are usually applied in conjunction with another function. They organise the records into alphabetical or numeric order, either ascending or descending.

Filtering

Filtering is powerful in database queries. Parameters are set to reveal only the records that are relevant. Any number of fields can be targeted so, for example, you could apply filters to a music catalogue database so that the only records that are displayed are songs by Elvis Presley released between the years 1968–1973 that are longer than 2½ minutes. These songs might then be sorted in alphabetical order by title.

Calculated fields

Calculated fields are newly created fields in queries that carry out some kind of manipulation. This might be a field that combines two existing fields, such as creating a new field *FullName* by combining *GivenName* and *Surname*. Or calculations may be required, such as calculating a student's age from their date of birth or a total price from the number of items multiplied by the item cost.

THINK ABOUT DATA ANALYTICS 2.5

Why do you think it is better to calculate something like age rather than have the user enter their age?

Designing queries

Queries need to be carefully planned. They can appear quite complex, but if approached logically they do not take very long to complete. They are best prepared using a table, in a similar way to designing the database tables. The most important thing is to understand what it is that you want the query to do. It might be a list of every member of a sporting group who has not paid their fees, of each person over a particular age, or to hide selected fields.

Table 2.20 is a sample query design. There is enough detail in this design for another person to create this without having to seek any clarification from the designer. Note that not all cells are filled in – only the cells that require action are completed. This query's purpose is to list all coins that are valid currency, in ascending order of value, which were acquired over five years ago.

TABLE 2.20 *qryOldValidCoinsOver50*

Field name	Sort	Filter	Calculation	Other
coiCoinID				Hidden
coiCountry				
coiDiameter				
coiMetal				
coiYear				
coiDateAcq				
coiValue	A–Z			
coiCurrency		>50		
coiShape				
coiValid		=True		
Age		>5	=(DateAcq-Now())/365.25	Hidden

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan | ☐ Collect complex data sets | ☐ Analysis | ☐ Folio of alternative designs | ☐ Infographic or dynamic data visualisations | ☐ Evaluation and assessment | ☐ Finalise report or visual plan

9780170440875

CASE STUDY

Schools in Australia

Part C: Creating queries

Julian and Carolyn specifically wanted data for two years – 2016 and 2017 – for the secondary schools in Victoria and New South Wales. This can be extracted by creating a single query, the design for which is shown in Table 2.21.

TABLE 2.21 Query design for school count case study

Field name	Sort	Filter	Calculation	Other
couID				Hidden
couYear		"2016" or "2017"		
couAgeLevel		"Secondary school"		
couSchType				
State	Ascending A–Z	"Vic." or "NSW"		
Count				

This query design does not need a calculated field. Look carefully to see if your SAC requires them (look for things like combined names and totals). Also note that any strings of text (all three filtered fields) must have the precisely correct content to work – no omitting spaces or changing punctuation.

Build the query as shown in Figure 2.7. Make sure anything in quotation marks is typed in accurately or else the query will not work as intended. Other operators that might be used in your SAC could be filtering values more than or less than a particular number, or finding a particular number. Use normal arithmetic devices such as > (greater than), >= (greater than or equal to), < (less than), <= (less than or equal to), = (equal to) or <> (not equal to) for these.

Field:	Year	AgeLevel	SchType	State	Count	ID
Table:	tblCount	tblCount	tblCount	tblCount	tblCount	tblCount
Sort:				Ascending		
Show:	☑	☑	☑	☑	☑	☐
Criteria:	"2016" Or "2017"	"Secondary school"		"Vic." Or "NSW"		

FIGURE 2.7 A query design in Microsoft Access

When run, the query produces a list of every record that meets the requirements, as shown in Figure 2.8.

qryVicNSW

Year	Lookup to tblLevel	Lookup to tblSchTy	Lookup to tblStat	Count
2016	Secondary school	Independent	NSW	15
2016	Secondary school	Catholic	NSW	127
2016	Secondary school	Government	NSW	369
2017	Secondary school	Independent	NSW	14
2017	Secondary school	Catholic	NSW	127
2017	Secondary school	Government	NSW	369
2016	Secondary school	Independent	Vic.	8
2016	Secondary school	Catholic	Vic.	86
2016	Secondary school	Government	Vic.	239
2017	Secondary school	Independent	Vic.	9
2017	Secondary school	Catholic	Vic.	85
2017	Secondary school	Government	Vic.	239

FIGURE 2.8 The results of the query built in Figure 2.7

9780170440875

Not all fields are needed for the user. In this case, the CoinID is not relevant – its only purpose is to be the primary key. It is not necessary for the user to see both the date of acquisition and the age, so age has been hidden from view.

Forms

Tables can appear intimidating, especially when full of data. They can be hard to read, and if a user is trying to find a particular record and read it across the screen, it can be tricky and lead to errors.

Forms are a user-friendly way for people to enter and view data in databases. Everything entered via a form is stored in the relevant database table. They are attractively and clearly laid out, displaying one record per view. This eliminates the chance of records being mixed up or misread. When entering data, it is the convention to work from the top left of a form and to navigate to the next cell by pressing the Tab key. Users find forms much less confusing and instinctively know how to fill them out since they resemble paper forms.

Designing forms

Form designs need to focus on both function and appearance. As such, the best tool to use is an **annotated diagram** since it can highlight any functional items such as buttons as well as indicate positions and formatting of objects.

Showing an example: the purpose of the form in Figure 2.9 is to allow the user to enter data about newly acquired coins.

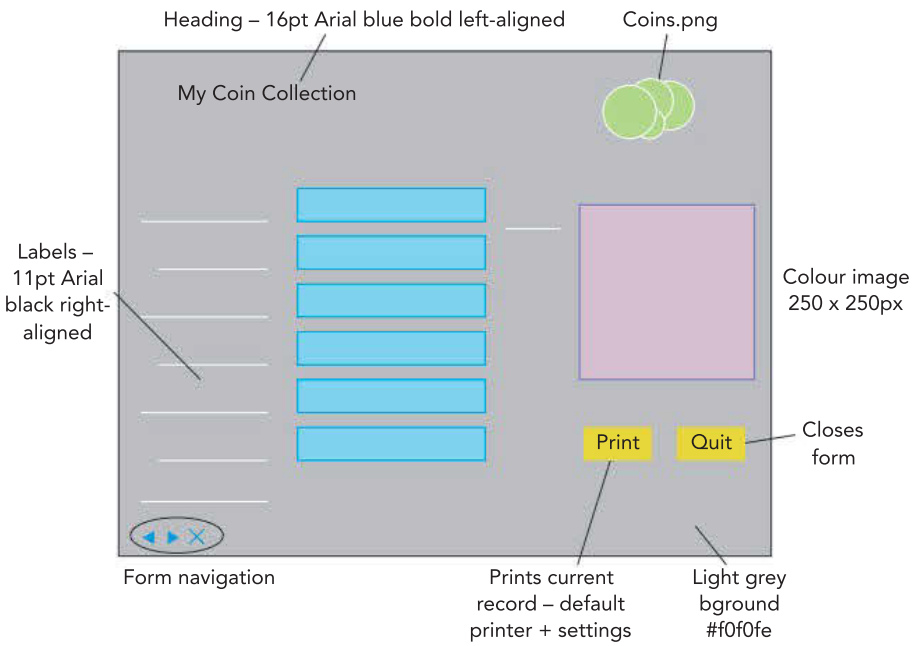


FIGURE 2.9 Annotated diagram showing appearance and functionality of an input form for a coin collection

Reports

When the database has done its job storing and manipulating data through the use of tables, forms and queries, the results need to be presented in a user friendly way. Just as forms make tables easier to understand by removing any extra information and being visually attractive, database **reports** present query results in an easily interpreted format.

9780170440875

Reports are usually presented as summaries of data. Only the relevant fields should be displayed (these may have been hidden when the query was designed, or they could be hidden when creating the report), and they should be presented in a way that is easy to read. Formatting choices here are important to improve clarity and readability, which in turn contribute directly to optimal communication of message for whoever the intended audience is. Often reports that are longer than one page will include automatically generated pages and final total figures if there are values to be added (such as an inventory value report).

Designing reports

Report designs are similar to form designs: they need to show the layout and appearance, but there are functionality issues that need to be addressed. If you are designing only one report, then you should indicate which fields will be required and any calculated fields that will be required (such as a running total cost for each page). If you are designing a generic report to be applied across many reports (basically creating a report template), then you cannot be overly specific. Good formatting and logical conventions should be followed, such as a heading at the top of the page, totals at the bottom, and a date of publication so that there is transparency around timeliness.

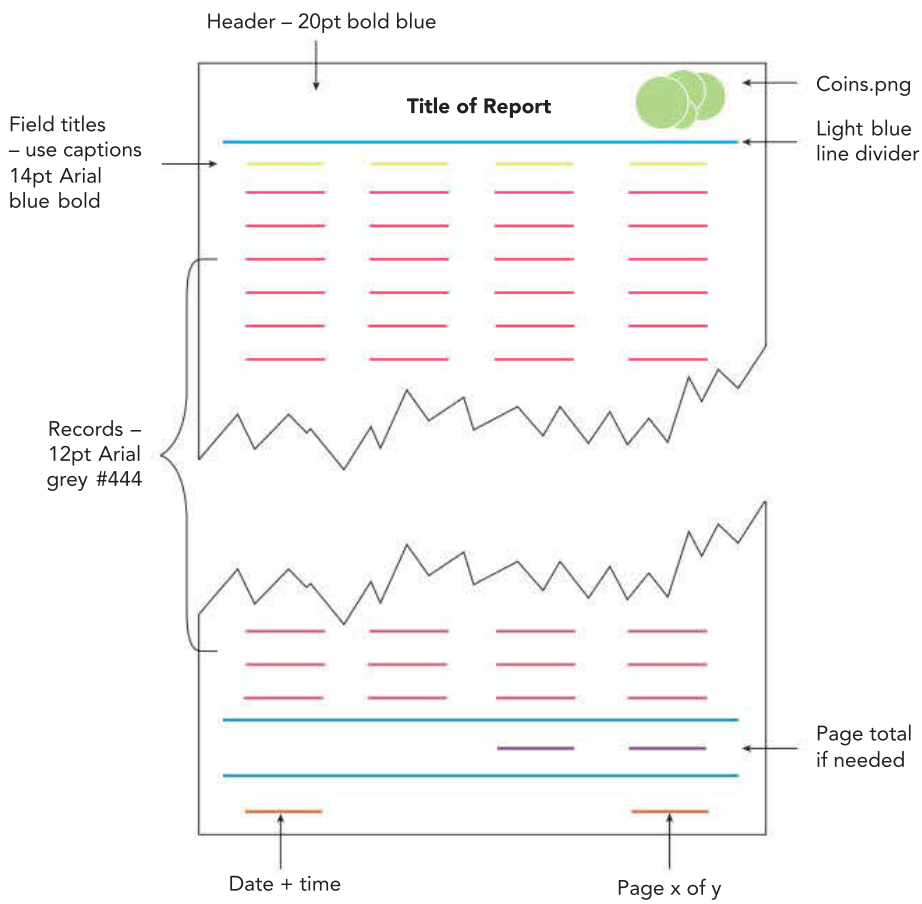


FIGURE 2.10 Annotated diagram showing placement and calculations required to build a generic report (**Note:** This could be applied to any report since no specifics are given)

Acquiring and inputting data

The next step is to acquire the data you will need. For your task, you may be gathering data first-hand (primary data) by using any of the methods described in Chapter 1: survey, interview or observation – although, survey is the only option for gathering large amounts of data quickly. Acquiring primary data yourself is time-consuming, and for this Outcome you are far more likely to use data that has already been collected. In your SAT, you will use both data collected by yourself and available data collected by others.

Manual entry

Typing in data manually is time-consuming and adds an extra step to the process. This is usually done when answers are being transcribed from a form that has been completed by hand (think about when you visit a new doctor or physical therapist). If the handwriting is unclear, or the person typing it up isn't paying attention, mistakes can easily happen at this stage. Validation rules can be set up to pick up some problems, but they would not be able to detect values that are incorrect – just values that are unreasonable, as discussed in Chapter 1.

Importing downloaded data

For large amounts of data collected by others, it makes sense to import it rather than type it in. It saves a lot of time, improving efficiency, and removes the chance of errors being made by misreading or mistyping values, thereby improving effectiveness.

When downloading data sets, it is important to understand that organisations and individuals supplying the data will want the files to be as small as possible to make them quickly and easily transferred. It is unlikely that you will find a large data set in the correct format for the database software you are using since these files are bigger due to the fact that they store more than just the data – they need to store information such as which sheets contain the data and formulas, and which also store the formatting.

Database software can also be used to acquire data from data sources. Using the external data function, data can be imported from a range of sources and file formats, including HTML documents, Excel, text and XML files, as well as other databases.

All spreadsheets and databases have the same error in their date calculations – they all have a February 29, 1900. 1900 was not a leap year; the error was made in the first database and has been replicated in all other computer programs so that the day numbers would be consistent.

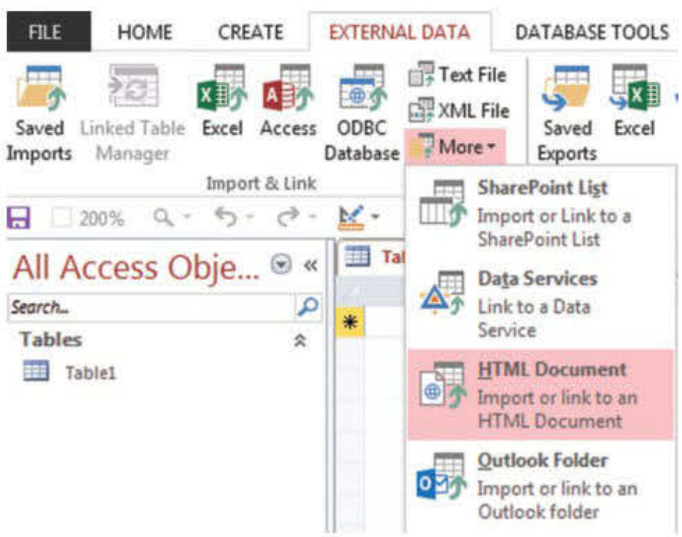


FIGURE 2.11 Options to import external data into database software

SCHOOL-ASSESSED TASK TRACKER

- ☐ Project plan
- ☐ Collect complex data sets
- ☐ Analysis
- ☐ Folio of alternative designs
- ☐ Infographic or dynamic data visualisations
- ☐ Evaluation and assessment
- ☐ Finalise report or visual plan

If importing data directly into a spreadsheet, the spreadsheet will interpret the data types for you based on the data in each cell. This can be problematic since sometimes the assumption the spreadsheet makes is incorrect. This happens often with entries that are meant to be numbers that are hyphenated or separated with slashes – spreadsheets will interpret these as dates. This becomes even more of an issue when the user attempts to convert the cell back to text since dates are stored as numbers where January 1 1900 was day 1, January 2 1900 was day 2, and so on (January 1 2020 is day 43 831).

This data type problem is precisely why we are importing our data into a database first. If you import the data without setting up a table, the database will assign data types according to what it thinks will work best. The problem with this is that if there are problems in the data, they will be ignored and any field with issues will most likely end up as a text field. However, if the fields are set up first, and the data types and validation rules have already been assigned, the validation will pick up any problematic data and quarantine it in a separate table for the user to manually check. This validation process saves a lot of time when dealing with large data sets.

Data cleansing

Data cleansing, also known as data scrubbing, is the process followed when data has been discovered (usually through validation) to be incomplete, inaccurate or lacking consistency. Microsoft Access will place all imported records that fail the data validation into a separate table. Once these records have been discovered, there are then several options that could be carried out, including:

- removing the entire record (deleting the whole entry, not just that piece of data)
- correcting the entry (if, for example, mistyped from a paper survey)
- modifying the entry to make it consistent with others (for example, changing gender entries so that they all read M, F or X instead of other combinations or spellings – M, m, male, Male)
- modifying the entry so that it complies with an expected format or convention
- modifying the entry to remove any additional information, such as moving a unit of measurement out of individual cells into a heading; this also makes it easier to perform numeric functions on the cells, because calculations cannot be performed on cells containing text.

Manually changing each incorrect entry is not an efficient option – it takes a lot of time, especially when dealing with big data.

CASE STUDY

Schools in Australia

Part D: Cleansing data

Once the data has been imported, check to see if there are any quarantined records or other alerts. Data acquired from sites such as the Australian Bureau of Statistics (ABS) often have excellent data integrity and validation breaches may not happen. If they do, manually examine the problematic data and make the choice to either modify it (be sure that it is an error, not something else you are changing), remove the record, remove the field but leave the rest of the record, or modify the validation rule to allow the value.

For your SAC, you will need to provide evidence of validation rule testing; run these tests before you import the real data.

The data from the ABS has extra information in it that we do not require. There are two efficient ways of removing this, depending on how your tables are set up.

9780170440875

Option 1

If the data is all in one flat file database table, use the 'Find and Replace' option. Make sure to use precise strings of characters: if you tell it to replace every 'a' character with nothing, then 'Tasmania' will become 'Tsmni'. Instead, instruct it to replace every instance of 'a' with a space after it. If you can see that this will still cause problems, replace 'a NSW' with 'NSW'. Always check dialogue box options carefully – in this case we have opted to replace the characters 'a' ('a' and 'a space'), but only if they occur at the start of the field. Also ensure the correct field is selected.

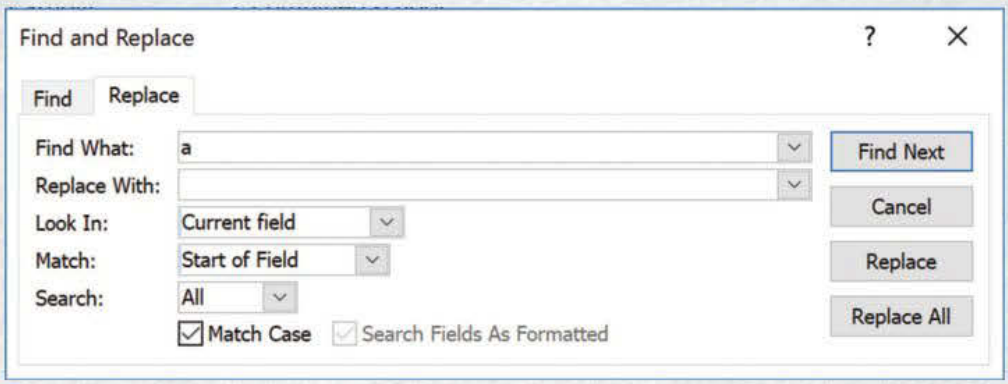


FIGURE 2.13 The Find and Replace dialogue box in Microsoft Access. The user-entered 'a' has a space after it, and the tool has been instructed to only look at the start of entries in the currently selected field.

h ACT
h ACT
h ACT
h ACT
h ACT
a NSW
a NSW
a NSW
a NSW
a NSW
a NSW
a NSW
a NSW
a NSW
a NSW
a NSW
a NSW
b Vic.
b Vic.
b Vic.

FIGURE 2.12 The imported data has unwanted additional characters at the start of each field entry.

b Secondary school
c Combined school
a Primary school
c Combined school
d Special school
a Primary school
b Secondary school
c Combined school
d Special school
a Primary school
a Primary school
b Secondary school
c Combined school
d Special school
a Primary school
b Secondary school
c Combined school
a Primary school
c Combined school

FIGURE 2.14 Again, the imported data has unwanted additional characters at the start of each field entry.

AgeLevel
a Primary school
b Secondary school
c Combined school
d Special school

FIGURE 2.15 The linked table is selected and the four options are changed, once for each.

Secondary school
Combined school
Primary school
Combined school
Special school
Primary school
Secondary school
Combined school
Special school
Primary school
Primary school
Secondary school
Combined school
Special school
Primary school
Secondary school
Combined school
Primary school
Combined school

FIGURE 2.16 The fixed fields automatically update in the first table.

9780170440875

Option 2

If your tables are linked, then repeated entries will be sourced from another table. In this case, each unique field entry will only need to be updated once. They will automatically update in the linked table. This shows the power of properly normalised data: each change is only made once and updates everywhere else.

Data that has been downloaded from reputable sources will most likely not need cleansing, but it is important to check – otherwise you cannot attest to the integrity of the data.

Common file formats for data sets

Files are a type of data structure that allows users to store and retrieve their data. Often, these are collections of data or data sets. Just as there are many types of data visualisations (as discussed in Chapter 1), there are many different ways to store the data sets from which they are created.

Here are some of the more common file formats for downloaded data sets. Most of these can be opened by spreadsheet software such as Microsoft Excel, Google Sheets and Apple's Numbers, as well as by databases.

CSV file format

A **comma-separated value (CSV) file** stores data in tabular format in a plain text file. The data is saved as individual fields using a comma to separate values. There are a number of advantages of saving data in a CSV format. First, a CSV is quite easy to import into (or export from) a range of software applications, including spreadsheet and database applications. Plain text files (such as that used by Notepad) also use very little storage space. Unlike some other file formats, they are saved as text, so they are readable.

```
1, Bunny, Bugs, Rabbit, 8
2, Pig, Porky, Pig, 3
3, Rabbit, Jessica, Human, 9
4, Brown, Charlie, Human, 7
5, Le Pew, Pepé, Skunk, 2
6, Martian, Marvin, Alien, 5
```

FIGURE 2.17 Comma-separated value file format

GIS file format

A Geographic Information System (GIS) file format contains geographical data. For example, it could contain the location of rivers and creeks in Victoria, stored as a series of longitude and latitude coordinates. For each river or creek, it could also contain attribute information, including the name and some history information. The GIS file will also contain data on how to display each creek and river in the visualisation.

XLM/XLMS file format

A file with the **XLM or XLSX file** extension is an Excel Microsoft Office Open XML format spreadsheet file. Microsoft Excel is the primary software program used to open and edit XLM and XLSX files. Alternatively, there are a range of free alternative spreadsheet programs that can access these files.

WMS file format

A **Web Map Service (WMS) file** is a file format that involves retrieving map images over the internet from a webserver. WMS files are used in conjunction with GIS file formats to create geospatial visualisations.

9780170440875



Copyright © The State of Victoria, Department of Environment, Land, Water & Planning 2015, Creative Commons Attribution 3.0 Australia

FIGURE 2.18 Web Map Service (WMS) using Geographic Information System (GIS) file data to show industrial areas of urban development. The areas outlined with purple lines are flagged as 'industrial nodes'.

Using queries to find patterns in data

Queries, as discussed earlier, can be used to find patterns in data. By setting up particular **filters**, you could pull out just the data that is relevant to your research topic. Sorting can also help you to quickly identify which items may require closer inspection. For example, if a population table indicated that few people lived in particular towns, you could filter out only those records to look more closely for anything else they had in common.

Exporting data

Once you have finished manipulating your data, it needs to be exported. The easiest and most versatile choice is to use CSV files. As discussed earlier, these are relatively small files that are able to be interpreted by other software tools.

Schools in Australia

CASE STUDY

Part E: Exporting database data

It is now time to export our data from the database software so that we can further manipulate it in a spreadsheet. Exporting is different to saving; saved files in databases can usually only be opened by the database software being used. Exporting involves changing the file type to something more readily transferrable. In our case, Carolyn and Julian have decided to export as a spreadsheet file because that will match the next software they will be using.

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

Spreadsheet tools

Spreadsheets will help you to manipulate data in order to develop graphs and charts. Spreadsheets can also accommodate small to massive data sets; include a large variety of charting tools and types; include chart style galleries that allow users to format each component of their graphic solution; allow users to enhance graphic representations, provide tools that enable users to highlight data trends; and present appealing, persuasive graphical summaries.

Spreadsheets are software tools that essentially perform calculations and create basic charts. Their core purpose is not about storing data efficiently – that is the database's strength.

Their true power is in the ability to perform complex mathematical functions. Mathematical functions include the ability to:

- perform basic arithmetic operations (+, –, /, *)
- perform statistical or other mathematical functions (average, minimum, maximum, median, standard deviation).

Spreadsheets can also perform complex logical functions, having the capacity to:

- create decision statements such as IF() or SUMIF()
- use a LOOKUP() table to extract data needed for a calculation from another worksheet
- use AND(), OR() or NOT() to create complex formulas.

Spreadsheets have the ability to produce different types of graphs and charts, having the capacity to:

- graph a series of data using a range of graph types (including bar and pie graphs)
- format a plot area and gridlines.

Spreadsheets have the ability to format data to meet the graphics needs of the user. Formatting functions include the ability to:

- insert labels (such as headings and subheadings)
- insert headers and footers (filename, page and date)
- insert notes and comments (to explain a function or provide help)
- insert borders and shading to add more meaning to the layout.

Spreadsheets are also used as a conduit to prepare data for importing into other software applications, which you will be doing both for Unit 3, Outcome 1 and your SAT.

When to use spreadsheets

Although they both deal with data and manipulation, spreadsheets are best suited when working with smaller amounts of data, performing quick calculations and simple sorting and analysis. Spreadsheets can also quickly and easily generate simple charts. More complex tasks or long-term storage of large data sets are better suited to databases.

For this Outcome, you will be acquiring a large set of data and manipulating and cleansing it in a database first before then working on it further with a spreadsheet and finally using data visualisation tools.

For more detail about these design tools, refer to Chapter 1 (page 26).

Designing spreadsheet components

It is best to make decisions about what you want the spreadsheet to look like before you begin working on it. The design tools commonly associated with spreadsheet construction are **input–process–output (IPO) charts**, **layout diagrams**, annotated diagrams and **mock-ups**.

9780170440875

Spreadsheet terminology

Spreadsheets, like most software applications, have their own terminology for specific things. This list is not exhaustive, and you should consider looking into some of the excellent online tutorials on using spreadsheets so that you get an understanding of how they work.

Cell

A **cell** is the smallest unit, one of the 'boxes' in a spreadsheet. Apart from headings, they should ideally only contain one piece of data – a name, number or calculation. They can be formatted in many ways.

Columns and rows

Columns are labelled by letters and contain all cells in a vertical line, while **rows** run horizontally and are referenced by a number.

Formulas

A formula in a spreadsheet performs some kind of calculation. It is preceded by an equals sign (=) to signify that a calculation is to occur. Table 2.22 shows some of the types of formulas that can be used.

TABLE 2.22 Examples of formulas in spreadsheets

Item	How to use it	Example
Basic arithmetic functions such as + − / *	=[cellref/number] operator [cellref/number]	=B8+17 =100-D7 =E2/C4*100

Functions

Functions in spreadsheets also perform calculations, but across a range of cells and use a word as a command. These cells can be listed individually, separated by commas, or they can be applied to a cell referenced range or named range.

Named ranges are a far more efficient way of applying functions. To name a range, simply select the cells and type in a logical name into the box in the top left corner that displays the cell reference.

TABLE 2.23 Some of the types of functions that can be used in spreadsheets

Item	How to use it	Example
SUM	=SUM (cellreftopleft:cellrefbottomright) =SUM (rangename) =SUM (n1,n2,n3....)	=SUM (C5:F17) =SUM (results) =SUM (B2,B8,C5,D8) =SUM (13,2,62)
AVERAGE	= AVERAGE (cellreftopleft:cellrefbottom right) =AVERAGE (rangename) =AVERAGE (n1,n2,n3....)	=AVERAGE (C5:F17) =AVERAGE (results) =AVERAGE (B2,B8,C5,D8) =AVERAGE (13,2,62)
MAX	=MAX (cellreftopleft:cellrefbottomright) =MAX (rangename) =MAX (n1,n2,n3....)	=MAX (C5:F17) =MAX (results) =MAX (B2,B8,C5,D8) =MAX (13,2,62)

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

Item	How to use it	Example
MIN	=MIN (cellreftopleft:cellrefbottomright) =MIN (rangename) =MIN (n1,n2,n3...)	= MIN (C5:F17) = MIN (results) = MIN (B2,B8,C5,D8) = MIN (13,2,62)
IF	=IF (condition to be met, true result, false result) Underlined sections may be nested (i.e. include another formula or an entire IF function)	=IF(B3 >60,"Yes","No") =IF(B3>$A$8,"Pass","Fail") =IF(B3>60,"Pass",B3) =IF(B3>60,"Terrific",IF(B3>30,"OK",'Drat'))
COUNT	Counts cells containing numbers in a given range	=COUNT (C5:E9) =COUNT (results)
COUNTIF	Counts number of cells with a stated value =COUNTIF (range,criterion)	=COUNTIF (C5:E9,"Pass") =COUNTIF (results,"B+")
VLOOKUP	Used to return a value from a specific column in a table =VLOOKUP (cellref/num,range,column)	=VLOOKUP (B3,$A$5:$E$12,2) =VLOOKUP (B3,results,2)
HLOOKUP	As above, but looks up a **row** instead of a **column**	=HLOOKUP (B3,$A$5:$E$12,2) =HLOOKUP (B3,results,2)
ROUND	=ROUND (cellref,numdecimalplaces)	=ROUND (4.5232,2) =ROUND (D4,3)

Relative referencing

When spreadsheets are given a cell reference (or cell address, such as B53), they do not 'look' directly at that particular cell. Rather, the program looks to see where that cell is located, relative to the cell where the formula or function has been entered. In other words, it looks to see how many cells up/down and left/right it needs to go to find what you require. This is called **relative referencing**, and it is very useful because it enables users to drag the calculation down a column (or across a row), and each time the cell reference will be relative. For example, if cell C1 was cell A1 divided by the contents of cell B1, and C1 was filled down, cell C2 would read =A2/B2 instead of =A1/B1.

Absolute referencing

Absolute referencing is used when it is necessary to always reference a specific cell, as opposed to wanting the ability to copy a formula's pattern. This is used, for example, if every number in a list needed to be multiplied by the same number. If that constant number was located in cell B4, then that cell would be what is called absolutely referenced. This is indicated in spreadsheets by using the dollar sign (an example of a program-specific convention), so in this case cell B4 would be entered as $B$4. Filling down with this would result in the other cell references changing, but the $B$4 would always appear the same.

Initial manipulation

Spreadsheets present CSV files as plain text, with one value entered into each cell and a new line begun with each new record. This means that when you open a CSV file in a spreadsheet such as Excel, the data will likely not be ready to use. Refer back to your designs to remind yourself of what you are trying to build.

At this point, this book will focus on how spreadsheets treat imported CSV files. If your data is in another format, some things may differ slightly. There are many guides on the internet to help you with this – just search for a tutorial.

9780170440875

The initial stage of working with data in a spreadsheet is to create headings as needed and move data around by selecting cells and cutting (Windows/Ctrl+X) to remove them and then pasting (Windows/Ctrl+V) into the new location. Most spreadsheet software tools follow common data conventions such as:

- numbers will align to the right of each cell
- text will align to the left of each cell
- any numbers that appear to be in date format will become date, even if that is not the intention.

CASE STUDY

Schools in Australia

Part F: Spreadsheet manipulation

Julian creates a chart using the spreadsheet and comments that New South Wales has a lot more schools than Victoria. Carolyn wonders what the population difference is between the two states, and if that has any impact on the relative numbers.

Carolyn goes back to the ABS website and finds the following information:

TABLE 2.24 Population of NSW and Victoria in 2016 and 2017

	2016	2017
NSW	7 480 228	7 861 068
Victoria	5 926 624	6 323 606

Source: ABS

They manipulate the spreadsheet by:

- removing the column stating they are all secondary schools (since it is redundant)
- pasting the state populations underneath the data
- adding a formula to account for the changes in population.

This is calculated by dividing the number of schools by the state population, then multiplying it by 7.5 million so that the NSW 2016 figures were the same. This formula is then applied to the other three state/year combinations so that they are comparable.

Note the use of both relative and absolute referencing in the formula. This formula refers to both the cell next to the one with the formula as well as one specific cell below (with the year's population). This reference was changed for each state and year.

F2 =D2/$D$16*7500000

	A	B	C	D	E	F
1	Year	State	School Type	Count		Adjusted Count
2	2016	NSW	Independent	15		15
3	2016	NSW	Catholic	127		127
4	2016	NSW	Government	369		370
5	2017	NSW	Independent	14		13
6	2017	NSW	Catholic	127		121
7	2017	NSW	Government	369		352
8	2016	Vic.	Independent	8		10
9	2016	Vic.	Catholic	86		109
10	2016	Vic.	Government	239		302
11	2017	Vic.	Independent	9		11
12	2017	Vic.	Catholic	85		101
13	2017	Vic.	Government	239		283
14						
15						
16		NSW	2016	7480228		
17			2017	7861068		
18		Vic.	2016	5926624		
19			2017	6323606		

FIGURE 2.19 A manipulated spreadsheet

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

THINK ABOUT DATA ANALYTICS 2.6

What impact would there be if no databases were capable of validating data?

Validation in spreadsheets

We covered validation in Chapter 1 and also earlier in this chapter when discussing databases. To review, remember that validation happens at the time that data is entered. It can only check that the data is reasonable; it cannot check that it is correct. Usually, validation occurs at the point of data entry, which was when we imported the data into our database.

There are three ways to validate data in spreadsheets: you can check for existence (that a cell is not left blank), the data type, and that data falls within a particular range. It is important to know that spreadsheets are more than capable of validation, particularly when data is being entered manually.

Electronic validation

When manually entering data, it is a good idea to set up cell validation prior to the data entry. Remember that validation checks the reasonableness of data – anything that restricts the data being entered to possible values is useful. Examples of this could include limiting cell entries to particular data types, setting acceptable ranges (for instance, only permitting temperatures between –10 and 50°C), or only allowing data from a fixed list (using a dropdown or combo box).

In this case, the **electronic validation** rules for the data can be set up after the data has been imported – select the cells and apply validation rules (in Excel, use Data > Data Validation). Excel can then apply those rules to any new data that is entered, or even circle the existing cells that do not meet the criteria.

Manual validation

The data should be checked (**manual validation**) for any noticeable issues the electronic process missed. Usually, this entails things such as spellchecking, looking for data entry errors, or anything else that seems out of place.

Formats and conventions in spreadsheets

Format and align data consistently. For example, ensure that all numbers in a specific column have the same number of decimal places and are right-aligned. This makes it far easier to see that 1000.0 is much higher than 1.00000 at a glance.

Appearance

Use colour and text formatting consistently; for example, shade all formula cells grey; use conditional formatting to turn cells red if their value dips below zero; and set headings as bold. Use 'Merge and Centre' to put headings above multiple columns or rows to unify them.

Units

Remove units from numbers and put them in column headings, such as 'Weight (kg)', rather than in the same cell as the number. Use a single consistent unit in a column. Convert mixed units (such as 2 minutes 30 seconds) to a common denominator (such as 150 seconds).

You would not put two data items in a single cell if you expect to process either data item. For example, '2 minutes 30 seconds' or 'small $1.00 large $1.50'. Use two cells to store costs of 'large' and 'small' items.

Use one unit of currency only. For example, do not put some prices in US dollars (USD) and others in Australian dollars (AUD).

9780170440875

Data visualisation

For the final part of your Outcome's development, you will be creating a data visualisation of the information that you have produced.

Review the types and purposes of data visualisations listed in Chapter 1 (pages 25–31).

Flowcharts can be used to show the procedure users follow to create a data visualisation.

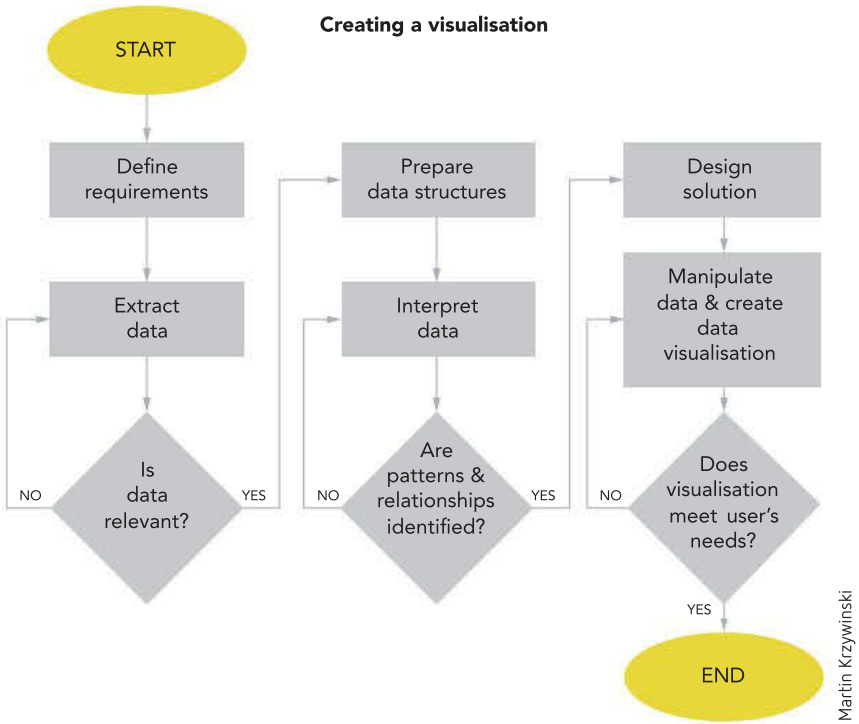


FIGURE 2.20 Flowchart on how to make a data visualisation

Designing data visualisations

You need to have a clear idea about what you want your visualisation to look like and how it functions. This includes any decisions that need to be made around the:

- type of visualisation
- appearance (colours, fonts, balance, use of space)
- interactivity (what the user will be able to control)
- multimedia components (moving images, sound).

The following are considerations and decisions that must be made before you can begin developing your data visualisation. You will not include every one of these elements – decide which best suit your needs and create those designs. Remember that you will need to test all of these and also design tests that can be used to check that each part works as intended.

Media and plug-ins

These include anything that is not text: images (pictures, photographs), audio clips, videos, interactive charts and animations. All of these need to be planned properly and carefully.

2.7 THINK ABOUT DATA ANALYTICS

How many data visualisations do you encounter on a daily basis? Would these have the same impact if they were presented as tables of raw data?

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

Think about how you would like your media to look and where it appears relative to other elements (see the annotated diagram examples for forms and reports in figures 2.9 and 2.10). When planning elements that involve video or animation, a storyboard is a good choice as your design tool, as shown below in Figure 2.21.

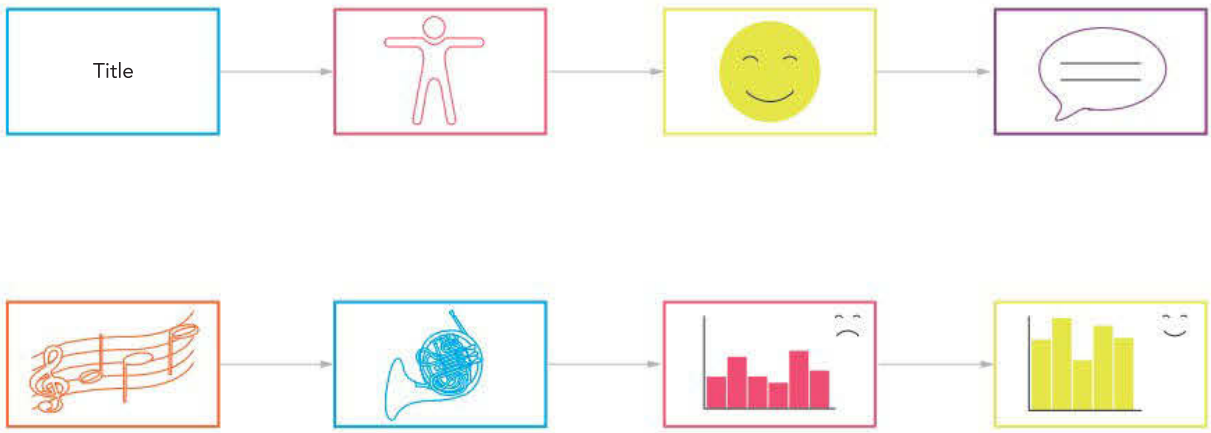


FIGURE 2.21 A sample storyboard sketching out the main ideas in a short animation

Hyperlinks

Any links in your solution, whether it be a website, interactive chart or something with clickable navigation, need to be planned during the design phase of the problem-solving methodology (PSM). Indicate any of this kind of functionality using annotations, which can provide the exact detail the developers need to create the correct product.

TABLE 2.25 Sample hyperlink testing table set-up during the design phase (completed during the testing stage in the development)

Item tested	How it was tested (data)	Expected result	Actual result	How it was fixed
Homepage *index.html*	Chart link Chart hover	Opens chart Colours change and alt text appears for each bar		
References page *refs.html*	Links to: ABS article CSIRO data	ABS article opens in new tab CSIRO data site opens in new tab		

Calculations

If your solution calculates any information, these calculations need to be planned properly. The best tool for this is an input–process–output chart (IPO chart). See tables 1.4–1.6 in Chapter 1 for a worked example of how to approach creating IPO charts.

Efficiency

Efficiency measures need to be considered when designing solutions to ensure that the solution saves the required time and that users can find and interpret the information with little effort. This ties in with the previous concerns: make sure that your solution is clearly thought out, with any links or interactive elements easy to find and use. These will need to be tested later on.

Many of these terminology elements overlap in regards to what they cover, and it can be difficult to separate them.

9780170440875

Effectiveness

Effectiveness measures also need to be planned for when designing solutions.

Completeness

Designs should incorporate all aspects of a solution – there should not be any omissions. Users should not have to look elsewhere for additional or supplementary information.

Readability

In your annotations, you should specify every aspect of your design that will enhance your work's readability. These include font choices and sizes, indicating colours to demonstrate high contrast and conventional alignment and placement of elements.

Text, audio and visual media all must be of a size, colour, volume, speed, illumination and fluency that can be interpreted with ease. The following section discusses two particular elements of readability: typefaces and white space.

- *Typefaces:* Use a plain, legible typeface for body text. Use either serif (such as Times New Roman) or sans-serif (such as Arial). Decorative, script and handwriting typefaces are not recommended. If you must use a decorative script or handwriting typeface, reserve it for headings only. Remember, you will not be presenting much text in your visualisation.
- *White space*: White space is a section of a graphic representation that is empty of any colour or object, which is used to create a clean, uncluttered look and is not considered wasted space by designers. In this instance, it refers to the empty parts of the screen that can be used to aid readability. The important areas of white space are margins at the edges of the page or screen, and visual gaps separating paragraphs and objects such as charts or headings (Figure 2.22).

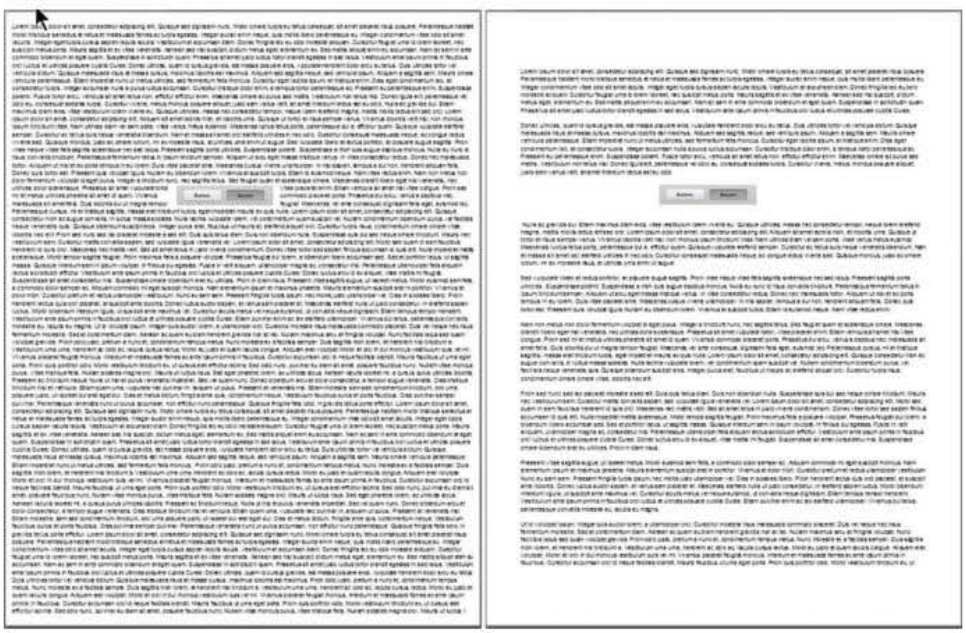


FIGURE 2.22 How white space affects readability. Poor use of white space is shown on the left while generous white space is shown on the right.

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

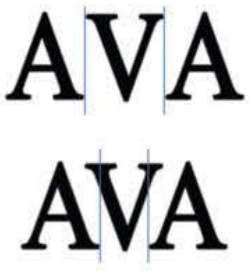


FIGURE 2.23 The top line has no kerning and the 'A' does not share any of the space occupied by the 'V'. With kerning, the A and V share space. The space between the lines is leading.

Refer to Chapter 1 (page 39) for more detail about conventions.

- *Leading:* Leading (pronounced 'ledding'), or line spacing, is the distance between two lines of type. Leading affects text legibility. It is set so the eye can flow easily from one line of text to the next. If leading is too tight, text is too close together. If leading is too loose, text is too far apart.
- *Kerning:* This refers to the spacing between pairs of letters. Certain letterforms need space between them adjusted when they are used together to avoid unsightly gaps, such as T and L.

Attractiveness

The interface is the only part of your solution that users will see. It must be appealing, attractive, and easy to use, regardless of how brilliant your information architecture is.

Consider the following for attractiveness of an interface.

- Give your solution a consistent look and behaviour.
- Make interfaces clean, simple, easy to learn, easy to use and attractive. Limit the numbers of colours and typefaces used.
- Avoid showing off. No-one will be impressed that you have discovered animated GIFs or have 200 typefaces to choose from – and that you are determined to use every one of them in your solution.
- Use subtle colours. Neon or overly bright colours will be difficult to tolerate for any length of time.
- Employ leading, kerning and white space to make text more readable.
- You should generally obey conventions. Doing things your own way just to be clever or different without a substantive reason will only annoy your audience.

Clarity

Whether spoken or written, your solution's language should be clear enough for most audience members to understand. The primary goal of any solution is to convey information to the audience. Make sure that you understand what your goals are as you produce each section.

Avoid technical terms and use direct language. This is especially true if you cannot avoid using **jargon**. Although it is a good idea to use short sentences and paragraphs, you need to keep the text or any explanations to a bare minimum.

Functionality

Design all aspects of the solution's functionality carefully. Once again, annotated diagrams allow the designer to add in a lot of extra information to ensure that the designs are precise enough for the developer to understand exactly what they need to create.

Accessibility

Accessibility refers to how your solution handles navigation and error tolerance. Your solution needs to have built into it the possibility that users may not understand how to use it – they may click the wrong area or use interactive elements incorrectly and alter things. Your solution should include ways for users to backtrack or reset things back to a default state.

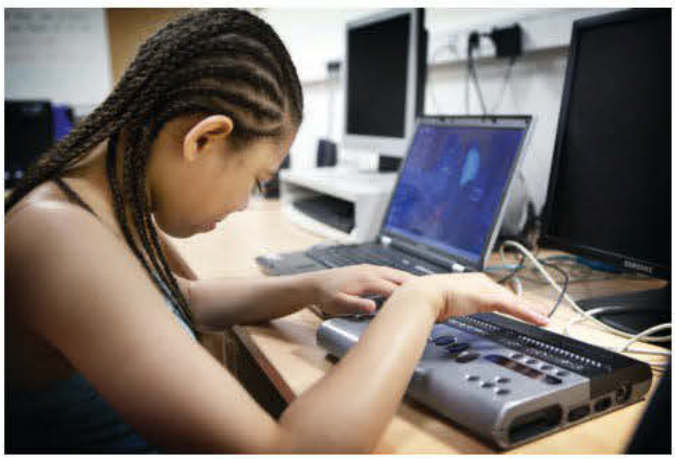
9780170440875

A poor interface is confusing, hard to learn and will cause user errors. It will discourage your audience from staying long enough to receive the message you aim to communicate.

- Anticipate common user errors and cope with them tolerantly, such as always providing a 'Back' or 'Cancel' facility so that users are not locked into a path from which they cannot escape.
- Grey-out any options that should not be selected.
- Be informative. Provide indications of how well a long operation is proceeding, such as how much has been downloaded or calculated so far, so that users do not think the computer is frozen or the website has crashed.
- Give help for tasks that might be a challenge, such as 'To download this video, right-click it and select "Save Link As".'
- Let typical end-users beta test your interface and pay attention to their feedback.
- Use the most appropriate graphical user interface (GUI) controls. For example, to force a choice of a single item from a selection, use radio buttons. To allow zero or multiple choices, use checkboxes.
- Be flexible. Your audience should be able to choose whether they use menus, shortcut keys or buttons to perform an action. Everyone has their own preference.

Another item that must be considered is the accessibility needs of users. For example, attach alt text to images so vision-impaired users with screen-readers can hear what the images represent. Avoid using red/green combinations and consider the end-user when choosing font sizes, colours and the size of any controls or buttons.

Alternate text (alt tags) are primarily used for images on web pages, but they can also be used for videos or interactive features. The tags provide a text-based description of the media's content. This will be displayed if the object fails to load in for some reason. They are also used by people with vision impairments; the text is read by the screen-reading software. If there is no alt tag, the reader will read the filename instead.



Getty Images / BSP / UIG

FIGURE 2.24 A vision-impaired user relies on the computer's speech synthesis to read the on-screen text out loud.

Timeliness

All data presented should be current. Outdated or old data may not be relevant or even useful in a presentation. There is no fixed amount of time that is considered acceptable since it depends on the context: 10-year-old data is useless when dealing with weather forecasts, but 10 years' discrepancy when mapping prehistoric data is very specific.

Another meaning of timeliness for the purposes of this task refers to the actual currency of your information. Refer to the 'Data integrity' section in Chapter 1 (page 20).

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

Displaying data

The following list shows tools and functions, other than Microsoft Excel, that can be used to display data.

Google

Google has a range of online cloud-based software tools that can manipulate data into a visual format. Google Sheets is a spreadsheet software, Google Charts offers a range of live streaming of data from websites, and Google Motion has the functionality to develop motion charts that can be used to create time visualisations.

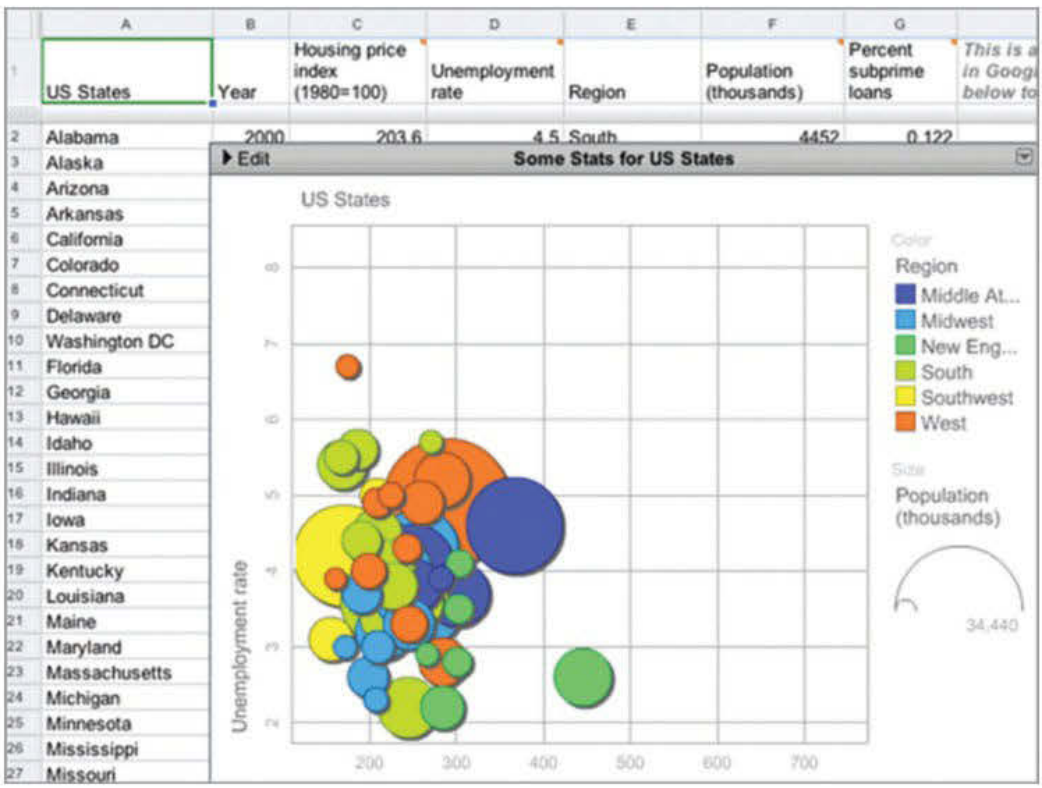


FIGURE 2.25 Motion chart created with Google Motion

Tableau Public
OpenHeatMap

Tableau Public

Tableau Public is a free software tool that can allow anyone to access a spreadsheet or file to create interactive data visualisations for the web.

OpenHeatMap

OpenHeatMap allows users to create static or animated heat maps. Data can be saved into files including Google Docs or spreadsheets and then uploaded to the site to create the map. Heat maps represent values in a range of colours that indicate concentration, similar to a weather map.



FIGURE 2.26 Tableau Public

Used with permission of Tableau Software

THINK ABOUT DATA ANALYTICS 2.8

The tools mentioned are merely a sample of those available online. Research other online tools and functions available to create data visualisations.

9780170440875

Infogram

Infogram has an assortment of visualisation and infographic generation tools. Free use is limited, but it is worth having a look at. There are word clouds, heat maps and a variety of chart formats available, many of which are interactive. The 'Schools in Australia' case study will make use of Infogram.

Infogram

Infogram is free to use for 10 data visualisations. Sign up for an account and investigate the vast array of charting and infographic options available.

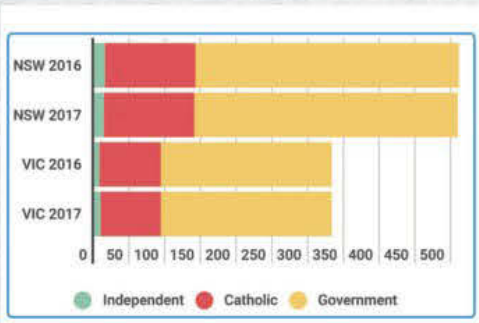
CASE STUDY

Schools in Australia

Part G: Creating data visualisations

For this project, Carolyn and Julian decided to use Infogram since they had both used it before. They entered the summary data and chose their chart type.

Figure 2.27 shows the data and chart generated using Infogram from manipulated data. Hovering over each section of the chart displays the data. The chart clearly shows that, relative to the state's population, New South Wales has more secondary schools than Victoria. It also shows that both states had fewer secondary schools in 2017 than they did in 2016.



	A	B	C	D	E
1		NSW 2016	NSW 2017	VIC 2016	VIC 2017
2	Independent	15	14	8	9
3	Catholic	127	127	86	85
4	Government	369	369	239	239
5					
6					
7					
8					
9					
10					
11					

FIGURE 2.27 Data and chart generated by Infogram using the manipulated data

Part H: Applying formats and conventions

Once happy with their chart choices, Carolyn and Julian need to add conventions. Changes that they choose to make include:

- changing the green to blue so that it is easier to distinguish from the red
- adding a title
- labelling both axes.

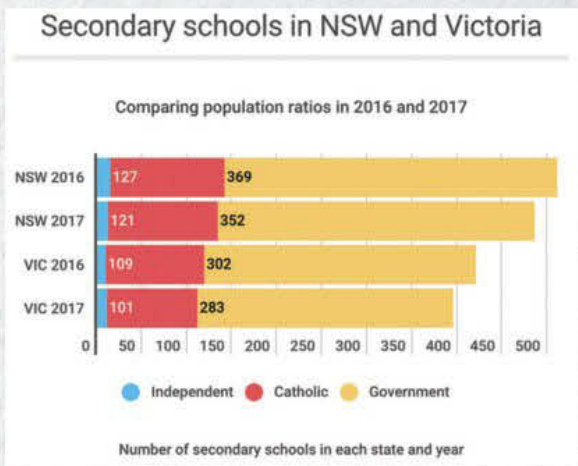


FIGURE 2.28 Completed visualisation

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

Testing

Make sure you understand the differences between testing, validation and evaluation, since they are not the same thing.

Testing is an essential step when creating any software solution. The purpose of testing is to ensure not only that the entire solution functions as intended, but also that each section of it does as well.

If a software solution fails in other contexts, not just in Data Analytics, it could annoy, disadvantage, or even kill, users, so thorough and careful testing is necessary whether the solution is a game, a website shopping cart, or an airliner's autopilot.

If your software solution fails because of undiscovered faults, it may become unpleasant to use, or completely unreadable.

Testing checks that a solution produces the correct output and does what it should do. Testing is not easy, quick or cheap – especially for a product such as an operating system with megabytes of code in thousands of files created by hundreds of people.

The typical steps involved in testing are as follows.

1. Decide which tests to conduct.
2. Create suitable test data.
3. Determine expected results.
4. Conduct the test.
5. Record the actual results.
6. Correct any errors.

There are many testing types, each intended to uncover different kinds of errors at different times during development. The types of testing relevant to your solution are listed in Table 2.26.

TABLE 2.26 Testing types

Type	What is tested?
Informal (alpha)	The part of the solution that has just been finished
User acceptance (beta)	Typical end-users use their own equipment to check that the finished solution is acceptable in real-world conditions
Component	A single part of a system works properly by itself (for example, a web form applies the correct delivery cost for a given destination)
Integration	Individual parts of a system work together (for example, the web form sends correct data to a separate database)
System	All components in the solution work properly as a single unit
Installation	The software is installed correctly and working on your system, server or domain
Compatibility	The software and its components are compatible with a variety of operation systems and browsers (if appropriate)
Usability	Whether users can operate the solution quickly and easily
Accessibility	Whether users with special needs or disabilities can use the solution

Main types of testing

Testing is a critical part of creating any kind of software solution. There are different categories of testing. Each is carried out at a different time in the development process and is usually performed by different people.

9780170440875

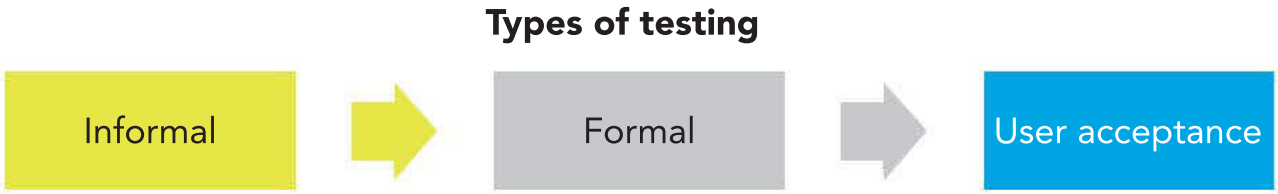


FIGURE 2.29 Three main types of testing

Informal

Informal testing is when you create an object or type in a formula and check it immediately to see if it works. When working in large teams, informal testing is undertaken by the developer and is not documented. There is no testing table or any kind of formal acknowledgement of the testing having occurred. This usually happens during the entire development process. For example, when a button is added to a form, the developer will usually click it to see if it performs as expected.

Formal

Formal testing usually occurs once the software solution has been completed, but remember that testing is still part of the development phase of the PSM. In large teams, this type of testing is carried out by testers. They use a testing table (see Table 2.27, page 100) that has the first three columns already completed, and **test data** identified. This would have occurred during the design stage of the PSM. Once any errors are identified, they are documented and the developers have the opportunity to make changes. This process may be repeated more than once.

A **bench test** forms part of the formal testing process in which sample data that has been developed during the **design stage** is used to determine how the solution behaves with a range of data. Data is chosen to see whether calculations are performed correctly and how erroneous data is handled by validation techniques. For each piece of sample test data, the expected outcome is determined and compared with the observed response of the solution. Table 2.15 showed an example of a **test plan** created during the design stage. This test plan is now used to allow testers to check each part of the solution to ensure its proper operation. The results of testing are normally written by hand on printouts from the solution.

Test data

To prove the accuracy of a solution's output, give it some test data to work on, and compare the solution's answer with one known to be correct.

Good test data includes the following.

- Valid data – data that is perfectly acceptable, reasonable and fit to be processed.
- Valid but unusual data – data that should not be rejected even though it seems odd. A 10-year-old might, once a century, enrol in university. Validation that rejected the young genius' enrolment would cause embarrassment.
- Invalid data – to test the code's validation routines. For example, if people must be 18 years old to be given a credit card, the test data should include people *under* 18 so they can be seen to be rejected.
- Boundary condition data – data that is on the borderline of some critical value where the behaviour of the code should change. These 'tipping point' errors are a frequent cause of logical errors.

9780170440875

Testing table validation

The first key feature to test in a database is table validation. Testing validation first ensures that the data is accurate before it is entered via forms, or queries are run. It is important in the testing process to choose data that will test the boundaries of the validation rule as well as the data type and even the existence of the data. For example, if the validation rule was set to check for dates in the past, appropriate test values would be to enter yesterday's date, today's date, and tomorrow's date. A simple rule of thumb is to test a value before the value itself and one value greater than the value in the validation rule. Figures 2.30–2.32 show examples of how to annotate testing.

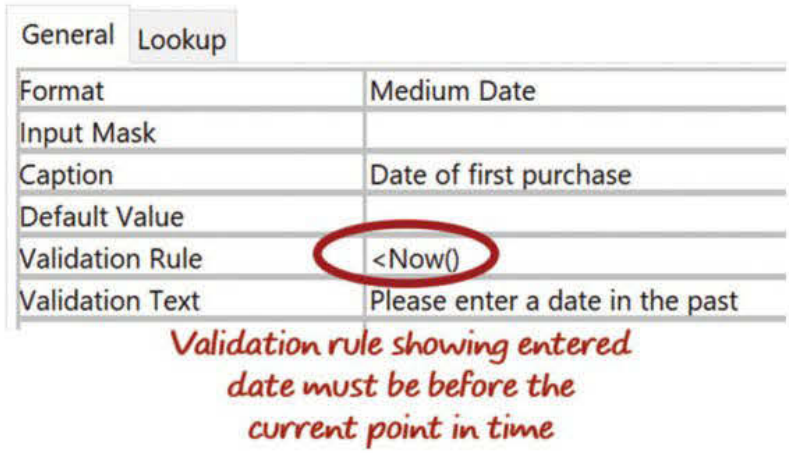


FIGURE 2.30 Screenshot showing validation rule preventing current or future dates being entered

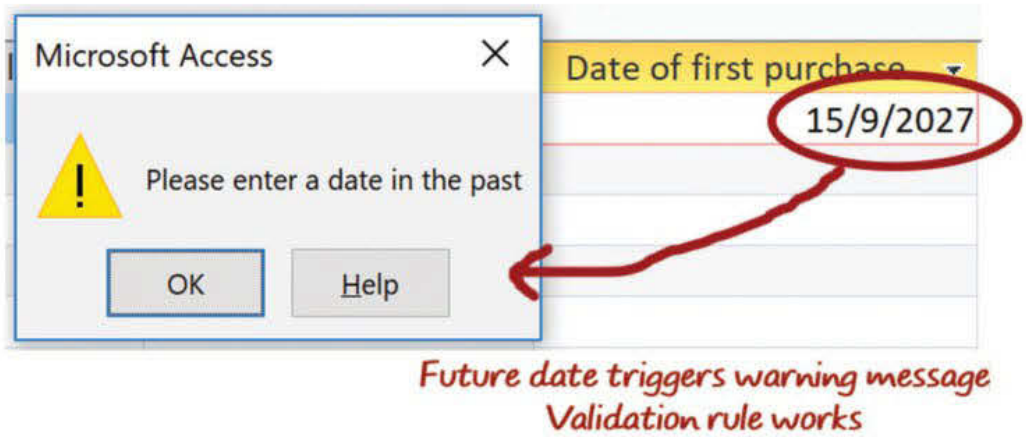


FIGURE 2.31 Test data of future date triggering a warning message

9780170440875

Testing the query selection criteria

Queries need to be tested to ensure that they select the correct data. Once you have created a set of test data, manually check that test data to identify the records that meet the criteria specified in the query. The test data must be chosen carefully to ensure that there are some records that should be returned by the query, some that should be rejected for failing to meet the first criterion even though they meet the second criterion, some that fail to meet the second criterion after meeting the first criterion, and finally, some that miss both criteria. You should note the records in the test data that should be returned and compare them against the records listed on the printout from the query. The records on both lists should match exactly.



FIGURE 2.32 Valid date accepted – user was able to click out, and the date entered in DD/MM/YYYY format was changed to the medium date format, adding an extra layer of validation because the user knows that the database has understood the date format.

qryVicNSW

Year	Lookup to tblLevel	Lookup to tblSchT	Lookup to tblStat	Count
2016	Secondary school	Catholic	ACT	5
2014	Secondary school	Catholic	ACT	5
2017	Secondary school	Catholic	ACT	6
2017	Secondary school	Catholic	NSW	127
2016	Secondary school	Catholic	NSW	127
2015	Secondary school	Catholic	NSW	129
2014	Secondary school	Catholic	NSW	130
2014	Secondary school	Catholic	Vic.	85
2017	Secondary school	Catholic	Vic.	85
2015	Secondary school	Catholic	Vic.	86
2016	Secondary school	Catholic	Vic.	86
2015	Secondary school	Government	ACT	19
2017	Secondary school	Government	ACT	19
2014	Secondary school	Government	ACT	19
2016	Secondary school	Government	ACT	19
2017	Secondary school	Government	NSW	369
2015	Secondary school	Government	NSW	369
2016	Secondary school	Government	NSW	369
2014	Secondary school	Government	NSW	370
2015	Secondary school	Government	Vic.	237
2014	Secondary school	Government	Vic.	238
2017	Secondary school	Government	Vic.	239
2016	Secondary school	Government	Vic.	239
2017	Secondary school	Independent	NSW	14
2015	Secondary school	Independent	NSW	14
2014	Secondary school	Independent	NSW	14

FIGURE 2.33 Extract of the full list of schools from the 'Schools in Australia' case study

Field:	Year	AgeLevel	SchType	State	Count
Table:	tblCount	tblCount	tblCount	tblCount	tblCount
Sort:					Ascending
Show:	☑	☑	☑	☑	☑
Criteria:	"2016" Or "2017"	"Secondary school"		"Vic." Or "NSW"	

FIGURE 2.34 Query set-up to only allow the fields that meet the specified criteria (for fields *Year*, *AgeLevel*, *State*) and in ascending order by *Count*

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

Testing sorting

A key feature of a query is being able to display its results in an organised and usable manner (that is, as information). These primary and secondary sorts have to be tested. The key is to demonstrate that a sort works alphabetically or numerically, or even alphanumerically. You will need at least three test values to demonstrate this adequately.

qryVicNSW

Year	Lookup to tblLevel	Lookup to tblSchTy	Lookup to tblStat	Count
2017	Secondary school	Independent	NSW	14
2016	Secondary school	Independent	NSW	15
2016	Secondary school	Catholic	NSW	127
2017	Secondary school	Catholic	NSW	127
2016	Secondary school	Government	NSW	369
2017	Secondary school	Government	NSW	369
2016	Secondary school	Independent	Vic.	8
2017	Secondary school	Independent	Vic.	9
2017	Secondary school	Catholic	Vic.	85
2016	Secondary school	Catholic	Vic.	86
2016	Secondary school	Government	Vic.	239
2017	Secondary school	Government	Vic.	239

FIGURE 2.35 The filters have worked as expected, but the sort did not. The error was fixed by going back into the query design and removing the sort on the *State* field.

Testing formulas and summary statistics

It is also critical to test formulas and summary statistics (such as 'counts' and 'sums') that occur in your queries. Choose appropriate test data so that it is easy to determine if the final result is accurate. The example shown in Figure 2.36 (page 103) can be easily adapted to demonstrate testing for a wide range of similar database or spreadsheet elements.

Elements than should be tested in databases and spreadsheets include:

- formulas (check that each formula works as expected)
- functions (check that each function such as SUM or COUNT works)
- conditional formatting (check that the correct cells are coded).

CASE STUDY

Schools in Australia

Part I: Testing

TABLE 2.27 Sample testing table (completed version of Table 2.15) showing one element from each of the database, spreadsheet and visualisation parts of the case study

Item tested	How it was tested (data)	Expected result	Actual result	How it was fixed
Database query: Only records with Vic., NSW, secondary schools and 2016–2017 are displayed	NSW, 2016, secondary Vic., 2017, secondary NT, 2017, secondary 2015 Primary	Accepted Accepted Rejected Rejected Rejected	All as expected	N/A
Spreadsheet: formula in cell F8 works	Manually check and compare	10	As expected	N/A
Data visualisation	Check chart for clarity	Easy to interpret	Difficult to work out what the message was	Switch axes around

9780170440875

User acceptance

The purpose of **user acceptance testing** is to ensure that the solution meets the actual needs of the intended users. It usually involves asking users to follow a series of steps to complete a task in the solution or interpreting the information contained in the output, then providing feedback to the developers. The testers are anticipated typical users of the system, such as when selecting a group of teenagers to test a PG-rated game. This is also called beta testing and occurs after all detected errors in formal testing have been corrected, but before the solution is considered finished. Any issues that are uncovered are reported and addressed, still as part of the development stage of the PSM.

Testing tools and documentation

A testing table is the most commonly used method to record evidence of functionality testing. A testing table for a website may look like Table 2.28 below.

TABLE 2.28 Sample testing table for a database query. Tests are run in late January 2020.

Item tested	How it was tested (data)	Expected result	Actual result	How it was fixed
ID field hidden	Run query, look to see if ID field is visible	ID field not visible	ID field visible	Uncheck the 'show field' box
A–Z sort on surname	Check each surname – they must be in alphabetical order	Alphabetical order	As expected	N/A
Calculated field *Age*	Enter DOB (date of birth): 24/1/2003 15/9/1990	17 29	6215 10 729	Divided result by 365.25 to convert from age in days to age in years
Filter to only include ages under 18	Enter ages: 17 18 19	Visible Not visible Not visible	Visible Visible Not visible	Change border condition to be <18 not <=18

How to document your testing

- Use a testing table such in Table 2.28 to simulate formal testing. Include several different types of testing and test different features. For example, demonstrate the testing of a range of items such as validation, formula correctness, positioning of elements, efficiency, elements of effectiveness, button or link operation, colour contrasts, and so on.
- Seek a subjective report from a fellow student who tried out your solution's readability and usability.
- Capture screenshots of features that are not normally visible, such as dropdown menus and warning messages, showing that they work when needed.
- Make handwritten calculations annotating printouts of screenshots of your solution's calculations to verify that the output has been checked for accuracy.
- Capture screenshots of the solution's validation rules responding properly to invalid data.

9780170440875

Testing data visualisations

After designing and building your data visualisation, you need to demonstrate that it has been thoroughly tested. The following items may or may not be a part of the solution you develop for this Outcome, but may be a part of the infographic you generate in your SAT. Test the elements that are included to ensure your solution appears and functions as intended.

Media and plug-ins

Each picture, audio clip, video, graph and animation (that is, any non-textual information), must be inspected to see that it is displaying in the right place, at the right time and at the right speed and volume in a variety of common environments (that is, different browsers and devices).

Hyperlinks

If links are included, every internal and external link in the solution needs to be manually clicked and the result noted. Create a list of links and tick each one off as it passes testing.

Readability

The checklist in Table 2.29 is a good guide to follow when checking for readability.

TABLE 2.29 Testing readability

Checklist	Tick ✓
Is the text large enough to read comfortably on a small device?	
Is contrast optimal, or at least satisfactory?	
Is the typeface a readable size?	
Are lines or paragraphs of a good length?	
Is text alignment attractive and readable on the page?	
Are the spelling, punctuation and grammar correct?	
Is the vocabulary appropriate and inoffensive?	
Is expression clear and unambiguous?	
Are headings clear and do they divide content into logical sections?	
Are all charts appropriately labelled?	

Calculations

If your solution calculates any information, its answers need to be verified by manual recalculation in a testing table. For example, you might insert a sum field to display the total number of survey respondents. To prove that you have tested the accuracy of its output, take a screenshot. Annotate the screenshot with whatever manual calculations will demonstrate that it is correct.

Loading times

If your visualisation is presented online, clear your browser's cache to remove pre-loaded copies of files and media and try loading the site via both wired and wireless connections. Any page that takes more than a few seconds to load should be inspected and optimised. Another method is to use one of many online services that can measure the loading times for your pages.

9780170440875

Browser compatibility

An online solution must rely on its browser, plug-ins and installed players and codecs (coder/decoder or compressor/decompressor) to read and display its media. Browsers differ in their ability to interpret different media, and some systems may not have the right technology, such as HTML5, or the necessary plug-ins installed, such as Adobe Flash Player or Microsoft Silverlight. Every piece of media must be checked on the dominant browsers to verify that they appear as they should.

You can test most site functionality yourself manually, but if your solution is online there are many services that can perform automated cross-browser compatibility checks using many versions of new and previous browser versions.

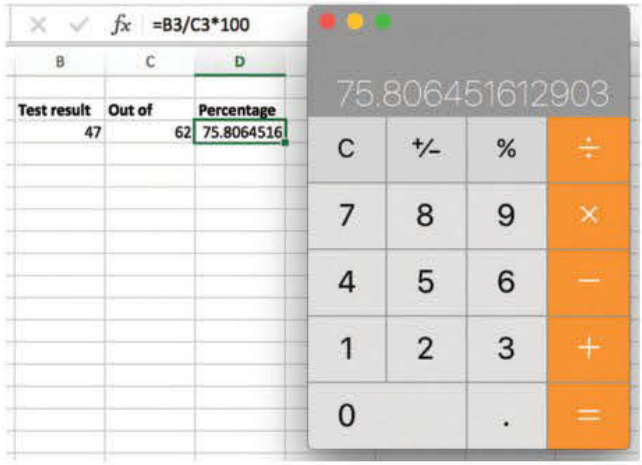


FIGURE 2.36 Testing that the formula worked in the spreadsheet by comparing the result to a manually calculated result

Dynamic features

Every menu item must be selected and its behaviour documented in a testing table (such as tables 2.15 and 2.27). If interactive elements are expected to work, data should be entered (if relevant) and its successful reaction should be documented. Any elements such as scrollbars, mouseover events or clickable items should be run and tested fully.

Efficiency

Efficiency measures need to be tested to ensure that the solution saves the required time and that users can find and interpret the information with little effort. This ties in perfectly with user acceptance testing and would only be conducted after the solution was completed for extra feedback to ensure that typical users can access the information quickly (time) and easily (effort). There is no need to directly test anything to do with monetary issues.

Many of these effectiveness measures will be investigated again during the evaluation stage of the PSM. It is important to check them during testing to pick up on anything that needs to be improved before the solution is considered fit for publishing.

Effectiveness

Effectiveness measures also need to be tested, and again they are well-suited to user acceptance testing. Elements of data visualisations that should be targeted for effectiveness testing include:

- completeness
- readability
- attractiveness
- clarity
- functionality
- accuracy
- accessibility
- timeliness
- relevance
- usability
- communication of message.

9780170440875

Completeness

Completeness means that the information you are presenting is just that: complete. You need to ensure that all of your information is presented without any omissions. You should refer to the 'Data integrity' section in Chapter 1 as a guide.

Possible sample questions for testing data visualisations:

- Are any significant points omitted?
- Are all headings and labels included?
- Are resources included?
- Is the data visualisation annotated when required?

Readability

It is important to test readability with typical users of the solution. All text must be able to be read with minimal effort.

Possible sample questions for testing data visualisations:

- Are the chart and its text of a readable size?
- Are colour choices appropriate?
- Is there enough white space to make the text uncluttered?
- Is there enough contrast to make the text easily read?

Attractiveness

The interface is the only part of your solution that users will see. It must be appealing, attractive, and easy to use, regardless of how brilliant your information architecture is.

Possible sample questions for testing data visualisations:

- Is the design visually pleasing?
- Is the visualisation balanced?
- Are the images used well and fit the colour palette?

THINK ABOUT DATA ANALYTICS 2.9

How could you test your solution for clarity?

Clarity

Clarity need to be tested by typical users to ensure that they have understood what has been presented to them.

Possible sample questions for testing data visualisations:

- Is there anything ambiguous on the visualisation?
- Is the visualisation easy to understand?

Functionality

Functionality will most likely be properly tested during the informal and formal testing stages. This will simply check that the solution operates as intended. This includes checking for browser incompatibilities.

Possible sample questions for testing data visualisations:

- Does the solution work with different devices?
- Does the solution make the software crash?

9780170440875

Accuracy

Accuracy refers to the precision of your visualisation. Your solution aims to educate your audience. Make sure that the information you provide is correct based on the data that you have carefully manipulated. Any data sources you used during Unit 3, Outcome 1 (and later, your SAT) should have been reputable.

Remember: Wikipedia itself is not an acceptable source of data. Any additional facts you cite must be checked in at least two places. Providing incorrect, outdated or misleading information is a violation of the trust an audience puts in an author.

Possible sample questions for testing data visualisations:

- Do the charts display the correct information?
- Are the charts/axes correctly labelled?

Accessibility

Navigation and error tolerance can be tested both formally and by typical users. You need to know that, if something goes wrong, the user will be able to fix it or go back to the beginning with minimal frustration.

Possible sample questions for testing data visualisations:

- Is the solution easy to navigate?
- If an error is made by the user, is it easy to undo it?
- Are 'Back' buttons provided where needed?

Timeliness

Your visualisation may be media-rich, and multimodal files can be large and slow to load. For it to be timely, your data should not be out of date. To determine this, you will need to put your data into context: if you are discussing the current weather, then a few hours old might be too old, but if presenting findings about geology, then data that is 10 000 years old may be perfectly acceptable.

Possible sample question for data visualisations:

- Is all of the information based on recent enough data?

Relevance

Regardless of your solution's topic, make sure you do not veer off onto another topic. Your audience is there for the stated message – your Outcome. They will have little tolerance for off-topic information. Only include material that relates directly to the research you are presenting.

Possible sample questions for testing data visualisations:

- Are all elements of the visualisation related to the topic?
- Is any of the information presented off-topic?

9780170440875

Usability

How user-friendly is your solution? Some of the user-friendliness aspects overlap slightly with the appearance design principles. However, you should also ask questions when creating, planning and testing your solution.

Possible sample questions for testing data visualisations:

- Is the user interface logical to use?
- Is it clear what the sliders/buttons do?

Communication of message

Communication of message ties all of the previous elements together. If any of the above are missing, your message will not be able to be understood and received as intended.

Possible sample questions for testing data visualisations:

- Does the user understand what the chart is summarising?
- Is the overall message clear?

Review the Outcome's steps

For Unit 3, Outcome 1, you will be presented with a problem. You will need to demonstrate skills in data cleansing, validation and manipulation using both spreadsheet and database software. After that, you will further manipulate the data using data visualisation software. Figure 2.37 gives you an outline of the process for the Outcome.

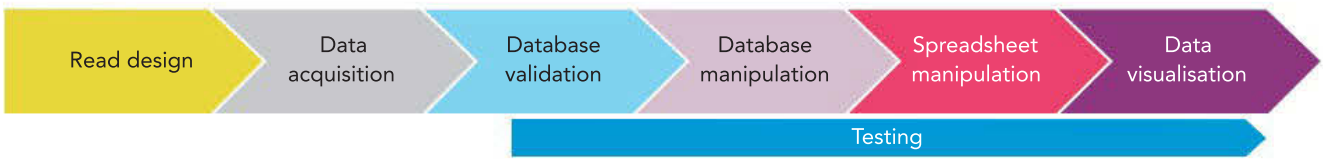


FIGURE 2.37 Outline of process for Unit 3, Outcome 1

2 CHAPTER SUMMARY

Essential terms

absolute referencing always referencing a specific cell in a spreadsheet that never changes

annotated diagram a rough sketch of the screen (form, report, web page) with written notes explaining the features in more detail

bench test a formal testing process to determine how the solution behaves with a range of data

calculated field a field in a database query that works out new fields based on existing fields, such as age from date of birth

cell where a column and row intersect in a spreadsheet; analogous to a field in a database

column a vertical series of cells in a spreadsheet

comma-separated value (CSV) file a file format that stores data in tabular format in a plain text file

data cleansing process where inaccuracies in data are detected and corrected either by changing, replacing or removing

data integrity ensuring that all data is trustworthy, which is achieved through accuracy, authenticity, correctness, reasonableness, relevance and timeliness

design stage a stage of the PSM; determining the way the problem will be solved

electronic validation a database's ability to detect limited types of errors in data that have been keyed, such as incorrect numbers entered for a postcode or certain boxes being empty, and then notify the data-entry operator of the error

entity a single person, place, or thing about which data can be stored

existence check validation that ensures a field has data entered into it

field an item of data in a database; the same field for a series of records will contain the same type of data

filter a process that only displays records that meet the set criteria

first normal form (1NF) a table with no repeating groups; that is, no single row has a column containing more than one value or more than one column with the same kind of value

flat file database a database that stores data in tables consisting of rows and columns; also called a single table database

foreign key a common linking field; a key defined in a second table that refers to a primary key in a first table (*see also* primary key)

form a window with input fields into which a user enters and then submits data for processing; forms are largely a replacement for paper-based forms and have features that include dropdowns, checkboxes, radio buttons, text fields and mandatory fields; common types of forms include email sign-up forms, online purchases, subscriptions and enrolments

input the process of entering data into a system

input mask a control applied to how data is entered that can be set to accept a specified number of digits, dates or characters in a commonly accepted format

input–process–output (IPO) chart diagram used to identify inputs, outputs and the processing steps required to transform the inputs into the outputs

jargon a shorthand way of communicating between members of a particular cohort

layout diagram a hand-drawn sketch that shows the elements to be included on an input form; the diagram would indicate the placement of fields and labels, the fonts to be used and any graphics or other elements to be included

manipulation where raw data is changed and processed to become information

manual validation the process of visually inspecting (proofreading) data to check if any mistakes have been made in copying the data; generally completed by a data-entry operator

9780170440875

many-to-many relationship when each record in the first table can be connected to multiple records in a second table

mock-up a sketch of a solution's appearance

naming conventions the way files, folders and database parts should be named for ease of use

normalisation removing redundant data and arranging it into appropriate tables to improve data integrity and reduce redundancy; involves following a systematic set of rules to check for anomalies or deviations in data structure, which ensures fields are in the correct tables

one-to-many relationship when a single record in the first table can be connected to multiple records in a second table (for example, several workers in an office with the same phone extension). The opposite is many-to-one

one-to-one relationship when a single record in the first table can be connected to only a single record in a second table (for example, a seat allocation table for airline passengers that holds records related to seats on a given flight, with a one-to-one relationship between each passenger and their allocated seat)

primary key a field attached to each record in a database; the value of this key should be unique for each record in the database (for example, numberplates on cars, tax file numbers and even your own student number); its purpose is to identify data related to each record, which may be stored across multiple tables

query to select specific data based on a series of criteria in order to answer questions and make links between data; the criteria are the results of questions about the data (for example, 'How many of our customers are female?')

range check validation rule to ensure that data falls within an acceptable limit

record a set of data about one entity (for example, a person, event or object)

redundancy repeated identical data entries

relational database a database that stores data tables that are arranged in rows and columns, with tables linked by a common field; relationships may be one-to-one, one-to-many or many-to-many

relationship connections between data

relative referencing cell references that change when copied across multiple cells

report the result of a query converted into a format that is usable and easier to read and understand that also allows you to add headings and summary statistics such as totals

row a horizontal series of cells in a spreadsheet

search finding only the relevant record(s)

second normal form (2NF) a table that is in 1NF and any column that is not part of the primary key is dependent on the whole primary key

sort arranging all records in order, either descending or ascending alphabetically or numerically

table a place in a database that holds data about a single person, place, or thing

test data a set of data that has been specifically formulated to reveal defects in software solutions

testing table a table set up to record functionality testing (what will be tested, how it will be tested, and what the expected result will be if the elements work as expected)

test plan a technique for recording tests to be carried out and the results of the tests; typically, a test plan states the type of test, what test data will be used, what results are expected and, ultimately, the actual results

third normal form (3NF) a table that is in 2NF and any column that is not part of the primary key is dependent only on the primary key and no other column

type check validation rule that checks that only the correct data type is accepted

user acceptance testing the last phase of software testing where it is determined whether the software functions according to specifications using real-world scenarios

validation checking that data input is reasonable, which is an activity within the development stage of the PSM; typically used for existence, range and type checking

Web Map Service (WMS) file file format that provides map images over the internet

XLM or XLSX file an Excel spreadsheet file

9780170440875

Important facts

1. There are many **file types** that can store large data sets such as .csv, .gis, .xlm, and .wms.
2. Common data types used in a **database** include text, numeric, date, character and Boolean.
3. A **field** contains the same type of data for a series of records. The same field for a series of records will contain the same type of data.
4. **Records** that meet specified criteria can be selected from the total number of records by performing a **query**. Each criterion can be a number, a piece of text or an expression.
5. The **structure for a database** can be determined by identifying entities, normalising tables, identifying fields and tables, identifying primary and foreign key fields, and defining data types and field sizes.
6. **Normalisation** is used in relational databases to reduce redundancy and make files smaller.
7. Use a consistent **naming convention** for database objects to make it clear what they are (for tables and queries) and what they are part of (for fields). Examples are *tblCustomers* for a table, or *qryAllProductsSoldToday* for a query. Fields should begin with letters that indicate the object they come from, such as *cusFamilyName*.
8. **Data cleansing** is used to check all data that fails validation tests so that the data is either fixed or removed from the data set.
9. **Data validation** at the point of input is essential to check for errors as the data enters the database or spreadsheet to prevent errors in the final product.
10. **Spreadsheets** are tools for performing calculations and generating simple charts.
11. **IPO charts** are used to plan calculations. IPO charts are used to identify inputs to a solution; how the data will be processed; and the required outputs of the solution.
12. **Data visualisations** can be created with a vast range of tools, many of which are available online.
13. Common **conventions** for data visualisations include a clear title, labelled axes, key or legend used, name of the author and source, units of measurement and appealing colours.
14. **Informal testing** is performed by the solution developer to ensure that components and functions are operating correctly.
15. **Formal testing** involves testing the overall solution after it has been completed.
16. **Test procedures** include establishing which tests will be conducted, determining which test data will be used, determining expected results, conducting the test, recording the results and correcting any identified errors.

9780170440875

TEST YOUR KNOWLEDGE

Review quiz

Databases

1 List three examples of naming conventions that could be used within a database.
2 How does a relational database differ from a flat file database?
3 Explain how a field and a record work together.
4 What is the purpose of a query?
5 Why are primary keys important in relational databases?
6 What makes a good primary key?
7 How does table normalisation help to ensure the integrity of data in an RDBMS?
8 What are some important features of a query design?
9 Explain the importance of establishing validation rules for RDBMS tables.

Spreadsheets

10 What are spreadsheets used for?
11 What is the difference between a function and a formula?
12 Why is it not a good idea to rely on spreadsheets for validating newly imported data?
13 List three different functions and explain their purposes.

Data visualisations

14 Identify a suitable type of data visualisation for the following data scenarios:
   a A breakdown of which nations are contributing to global warming by percentage
   b A comparison of three different aspects of national health – town population, average income and education levels
   c How to display the chemical elements in the human body

Testing

15 Why is it necessary to select appropriate test data during the design stage?
16 During which stage of the problem-solving methodology does testing take place?
17 Why is testing performed during this stage?
18 What should be documented during formal testing?
19 What is the purpose of user acceptance testing?
20 Why is testing important?

9780170440875



APPLY YOUR KNOWLEDGE

1. Explain the purpose of data normalisation. Make reference to the first, second and third normal forms.
2. Explain the differences between validation and testing.
3. Explain why it is a good idea to validate data in a database and not a spreadsheet.
4. Find a data set that describes Australia's population.
   a. Calculate each state's proportion of Australia's population and display as a data visualisation.
   b. All states have 12 senators in the Australian Senate, while the Australian Capital Territory and the Northern Territory have two. Calculate the number of people in each jurisdiction for each senator. Create a data visualisation of the results.
5. Using a heat map application online, create a data visualisation that shows which countries have the greatest populations relative to their geographic size.

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

PREPARING FOR Unit 3

OUTCOME 1

You will be given a scenario and a design brief that will outline the analysis of a problem.

Skills you will need to demonstrate

1. Interpreting designs
2. Choosing appropriate data
3. Extracting relevant data
4. Use of appropriate validation
5. Manipulating and cleansing data
6. Using database software
7. Using spreadsheet software
8. Selection, use and justification of tools to create data visualisations
9. Application of testing techniques

Steps to follow

1. Carefully read the design briefs and identify the solution that needs to be produced.
2. Identify the information that you need to create. The queries that you produce will be based on the information needs of the organisation. What needs to be communicated and to whom?
3. Create a testing table, identifying test data and expected results for database and spreadsheet.
4. Locate a suitable data set.
5. Create database tables according to the provided designs.
6. Import the data set.
7. Cleanse the data, removing or modifying records to ensure all data has integrity.
8. Create queries to extract the required data.
9. Complete the testing table database section.
10. Export the data from the database.
11. Import the data into a spreadsheet.
12. Further refine the data by adding relevant functions.
13. Complete the testing table spreadsheet section.
14. Export the refined data.
15. Import the data into an appropriate data visualisation tool to produce visualisations, ensuring that appropriate formats and conventions are followed.
16. Justify choices made around the selection of functions, formats and conventions.

This task is marked out of 100 and is worth 10% of your study score.

9780170440875

CHAPTER

3 Project management and data analysis

KEY KNOWLEDGE

After completing this chapter, you will be able to demonstrate knowledge of:

Digital systems

- roles, functions and characteristics of digital system components
- physical and software security controls used by organisations for protecting stored and communicated data

Data and information

- primary and secondary data sources and methods of collecting data, including interviews, observation, querying of data stored in large repositories and surveys
- techniques for searching, browsing and downloading data sets
- suitability of quantitative and qualitative data for manipulation
- characteristics of data types and data structures relevant to selected software tools
- methods for referencing secondary sources, including the APA referencing system
- criteria to check the integrity of data, including accuracy, authenticity, correctness, reasonableness and timeliness
- techniques for coding qualitative data to support manipulation

Approaches to problem solving

- features of a research question including a statement identifying the research question as an information problem
- features of project management using Gantt charts, including the identification and sequencing of tasks, time allocation, dependencies, milestones and the critical path

Interactions and impact

- key legal requirements for storage and communication of data and information, including and human rights requirements intellectual property and privacy.

Reproduced from the VCE Applied Computing Study Design (2020–2023) © VCAA; used with permission.

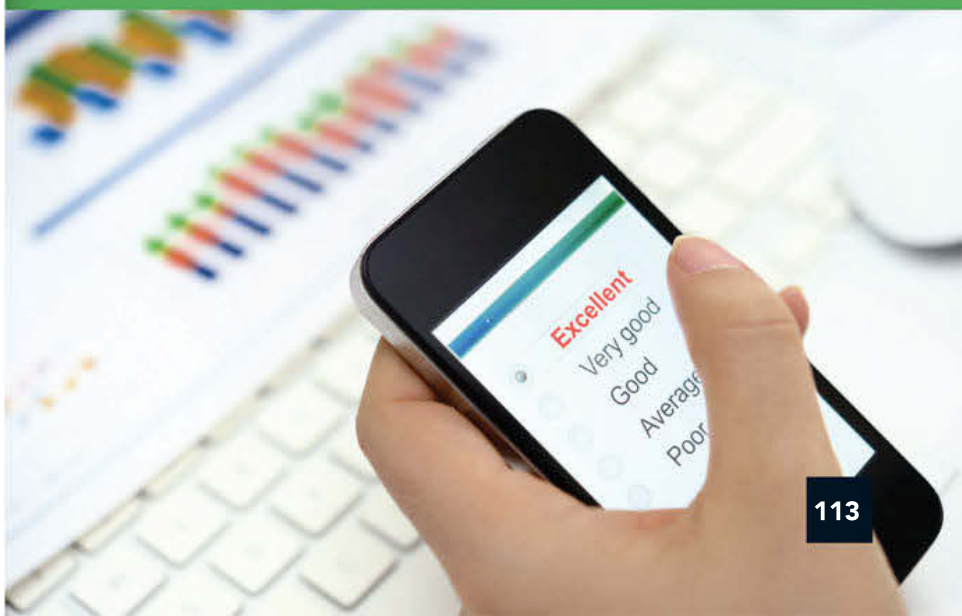
FOR THE STUDENT

This chapter begins the discussion of the theory and skills required for Unit 3, Outcome 2, which is the first part of the School-assessed Task (SAT) that will be completed in Unit 4, Outcome 1. Throughout both Unit 3 and Unit 4, you will be required to maintain a project management progress report.

The SAT requires that the project management report be submitted twice. In Unit 3, it will be in the form of a proposed plan, and in Unit 4 it is an amended plan and, with the benefit of hindsight, includes changes that reflect the actual project progress, rather than the imagined progress as recorded by the proposed plan in Unit 3. In this chapter, you will also be introduced to forming a research topic or question, legal requirements, types and sources of data, methods of data acquisition, methods of referencing sources, and how to check the quality of data.

FOR THE TEACHER

This chapter, along with Chapter 4, covers the theory required for Unit 3, Outcome 2. This is the first part of the SAT. The second part of the SAT will be completed in Unit 4, Outcome 1. Chapter 3 focuses on data: its types, acquisition, referencing, integrity, coding and the legislation relevant to storing and communicating personal data.



iStock.com/alubalish

What is data?

In Chapter 1 you learnt about data. In this chapter, you will learn about research and data.

To prepare you for Unit 3, Outcome 2 and also for Unit 4, Outcome 1, we will discuss planning and managing a complex project, including how to use a Gantt chart, because this relates to the project you are undertaking.

As part of Unit 3, Outcome 2, you need to formulate a research question, collect **data** and then analyse that data. The first thing we will discuss is what makes up a reasonable research question and how to formulate one for your Outcome.

Next, we will talk about how to determine the specifications for a solution, both in terms of general solutions and your specific solution, to help you understand how to determine factors such as scope and constraints. Then, we will discuss the data you need to collect, in terms of how it is categorised when it is interpreted by you when you are the main researcher, and how it is categorised when it is interpreted by someone else. Then, we will cover how data can be categorised in a different way – as either quantitative or qualitative.

The first thing to do after choosing a research question is to gather data, so we will cover how you can acquire data through sources such as libraries and data sets, and then through methods such as interviews and observation. We will also discuss how to reference those sources properly. Chapter 1 also covered some of this material, so make sure you refer back to it.

Once you have gathered all of this data, you need to store it, protect it and understand what type of data it is. We will discuss integrity of data and how to maintain it, through measures such as timeliness, accuracy, authenticity and relevance. This is important because you need to maintain the integrity of the data you collect for your Outcome so that your results and findings are considered reliable.

Finally, we will briefly cover data types and structures, because this is relevant to your Outcome.



FIGURE 3.1 Chapter map

Project management

Project management is the process of planning, organising and monitoring a project to complete it on time and within budget. Building or changing information systems for a project can be expensive and disruptive. If managed badly, it can be damaging to an organisation's operations and profit. Large-scale changes are often approached as projects so they can be planned, organised and conducted appropriately to enable them to finish on time, on budget and fully functional. You will formulate a project plan to manage your progress through Unit 3, Outcome 2 and in Unit 4, Outcome 1.

For your project to be successful, you need to identify, schedule and monitor tasks, resources, people and time. While you can use a software tool for planning a project, our main focus initially will be on the **concepts** and **processes** of project management.

One of the items you are required to submit as part of your Outcome is a **Gantt chart**.

9780170440875

A Gantt chart is a graphic timeline that:

- lists all tasks in a project
- organises the tasks in order
- shows which tasks must wait for other tasks to finish before they can begin
- allocates people and resources to tasks
- tracks the progress of tasks throughout the entire project.

Although you can create a Gantt chart with a pen or spreadsheet, project management software is usually easier and faster. Suitable software includes the commercial Microsoft Project, and the free, easy-to-use GanttProject and ProjectLibre.

When using software to create your Gantt chart, you will not be assessed on your technical prowess with the software. Rather, you will be assessed on how well your Gantt chart demonstrates your understanding of the concepts and processes of project management (Figure 3.2).

Students are not required to purchase commercial Gantt chart software to complete the School-assessed Task (SAT). Freely available open source software and templates are adequate and sufficient to satisfy all assessment criteria.

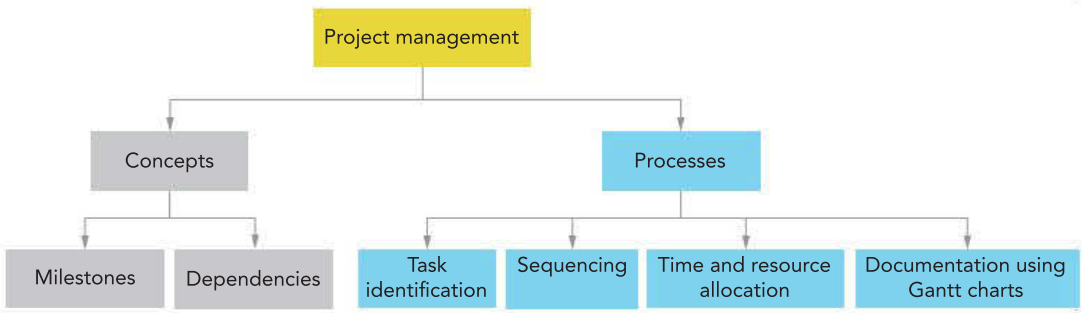


FIGURE 3.2 Key project management concepts and processes

Concepts

Milestones

A milestone represents the achievement of a significant stage in a project and has zero time duration. For example, the completion of the printing of a questionnaire so that it can be distributed to respondents would be a task of zero time and represents a milestone. This follows tasks in which the questionnaire has to be researched, written, proofread and, finally, printed – all of which take time. Milestones are usually indicated on Gantt charts by a diamond-shaped symbol.

Dependencies

Tasks are interdependent, meaning that they must be completed in a particular order. The commencement of some tasks depends directly on the task that is completed before. For example, you cannot distribute a questionnaire (one task) before writing the questions for it (another task). However, you cannot write the questions for the questionnaire without first determining your research question (a third task). Ultimately, the task of distributing a questionnaire has multiple dependencies. To distribute the questionnaire without first formulating the research question and writing the questions is not possible.

9780170440875

Processes

Task identification

A large project like Unit 3, Outcome 2 can be broken down into discrete tasks such as the following.

- Research multiple topics.
- Create a Gantt chart to record tasks for Unit 3 and 4.
- Formulate research question.
- Write questions, then conduct interviews and distribute questionnaires.
- Collate, input and interpret primary and secondary data.
- Conclude whether the research question has been supported.

Note: Not all of the tasks you would undertake for Unit 3, Outcome 2 are included in the list above. You should not use this as an exhaustive list.

While the WBS is not a part of the required SAT report, it will assist in assembling details and tasks for your Gantt chart.

To break down your project into achievable tasks, develop a **work breakdown structure (WBS)** and draw a WBS diagram to accompany it. Do not leave any tasks out of the WBS. For large projects, a WBS will often be hierarchical, breaking major tasks into subtasks and even sub-subtasks. Although this may sound confusing, it will actually keep your tasks organised and in context, allowing them to be collapsed or expanded to view overall task progress or fine details about how minor subtasks are proceeding.

Do not leave any tasks out of the WBS. For example, imagine you distributed both a print and online version of a questionnaire, but did not list the printed version in the WBS, and forgot to collect the printed forms from respondents. All the gathered data would be overlooked or would be counted too late.

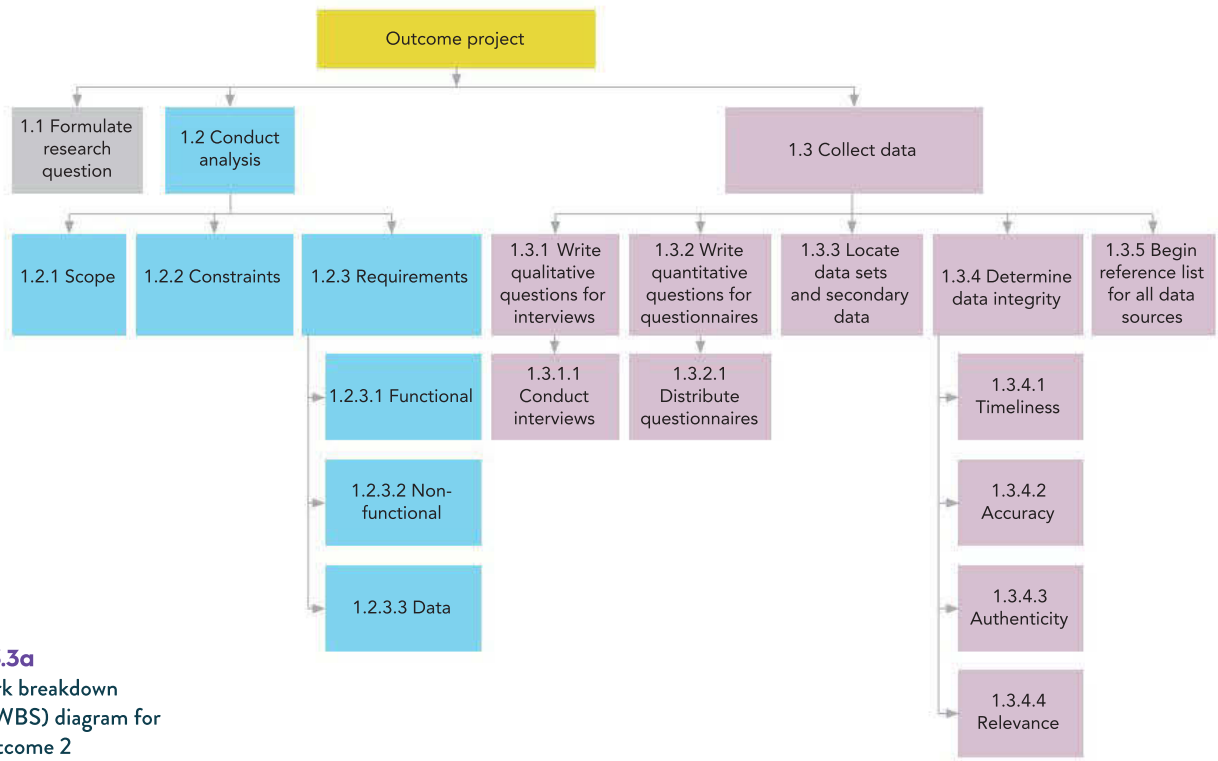


FIGURE 3.3a
Sample work breakdown structure (WBS) diagram for Unit 3, Outcome 2

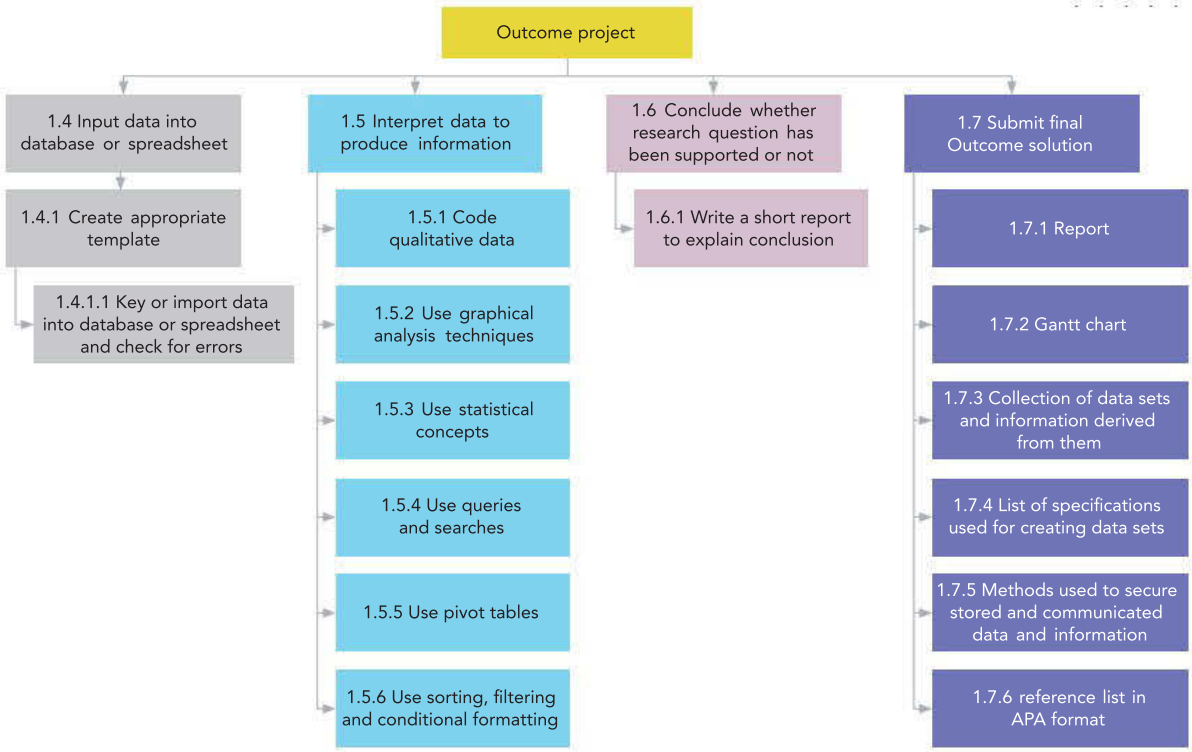


FIGURE 3.3b Sample work breakdown structure (WBS) diagram for Unit 3, Outcome 2 (continued)

Figures 3.3a and 3.3b show a sample potential WBS for Unit 3, Outcome 2. Again, the tasks may not be exhaustive and you may find that your own WBS requires additional tasks, particularly if you use different types of data, or you need to revise your research question.

Sequencing

When you have finished identifying every individual task, you need to decide how long each task will take and then put them in a sequence – that is, arrange them in a particular order. As discussed under the 'Dependencies' heading, quite often one task cannot be started before one or more other tasks have already been completed.

Decide what you can work on as **concurrent tasks** but are dependent on other tasks that have been completed already. For example, you could work concurrently on conducting interviews, researching data sets and writing quantitative questions for a questionnaire, but they are all dependent on the research question having been formulated first.

Similarly, you cannot write the report concluding whether your research question has been supported or refuted before you have interpreted your data.

Tasks that must be completed before another task can begin are called **predecessors**.

The dependent tasks are called **successors**.

If a predecessor runs overtime, all of its successors will be delayed, causing problems for other tasks and deadlines. This is where a Gantt chart becomes very useful – it helps to monitor tasks and reach deadlines on schedule. It also helps you to visualise the problems that will occur down the line if a predecessor is late.

9780170440875

The length of time that a task runs overtime before it affects other tasks is called **slack time**. When workers have slack time, you can reassign them to other tasks.

Time and resource allocation

A Gantt chart shows tasks as horizontal bars. Each horizontal bar is of a length proportional to its task's duration. A short task will have a short bar, while a long task will have a long bar. Figure 3.4 displays a number of features typical to a Gantt chart.

The names of each task are shown in the left pane, along with start and end dates, while the right pane shows task timelines. Tasks that overlap in time are concurrent and can be carried out at the same time using different teams.

Arrows are used to indicate dependency. For example, neither 'Write qualitative questions for interviews' nor 'Research data sets' can begin until 'Formulate research topic/question' has finished because they depend on this task.

The diamond shape indicates a milestone. Milestones are points of significant progress in a project. They are often the start or end of major stages, and they can be used to monitor whether a project is on track. A milestone is an **event** with zero duration and no allocated resources. It is simply shown as a diamond-shaped 'task'. In this instance, 'Submit final report', the end of the project, is one such milestone.

An event differs from a task because though something happens (for example, a major task ends), no resources, work or time are allocated for it because there is nothing that people need to do to make it happen.

Also notice that 'Input collected data' begins before both 'Research data sets' and 'Distribute and circulate quantitative questionnaires' have finished, even though the task is dependent on both of these predecessors. This is because 'Conduct interviews' *has* finished, so you can get a head start by inputting the data from the interviews while continuing to look at data sets and waiting for responses to questionnaires.

A project's critical path is the sequence of tasks from the beginning to the end, which:

- contains no slack time (therefore any delay in a task on the critical path will affect the ending date of the project)
- is the longest duration
- is the minimum possible time in which all of the project's tasks can be completed.

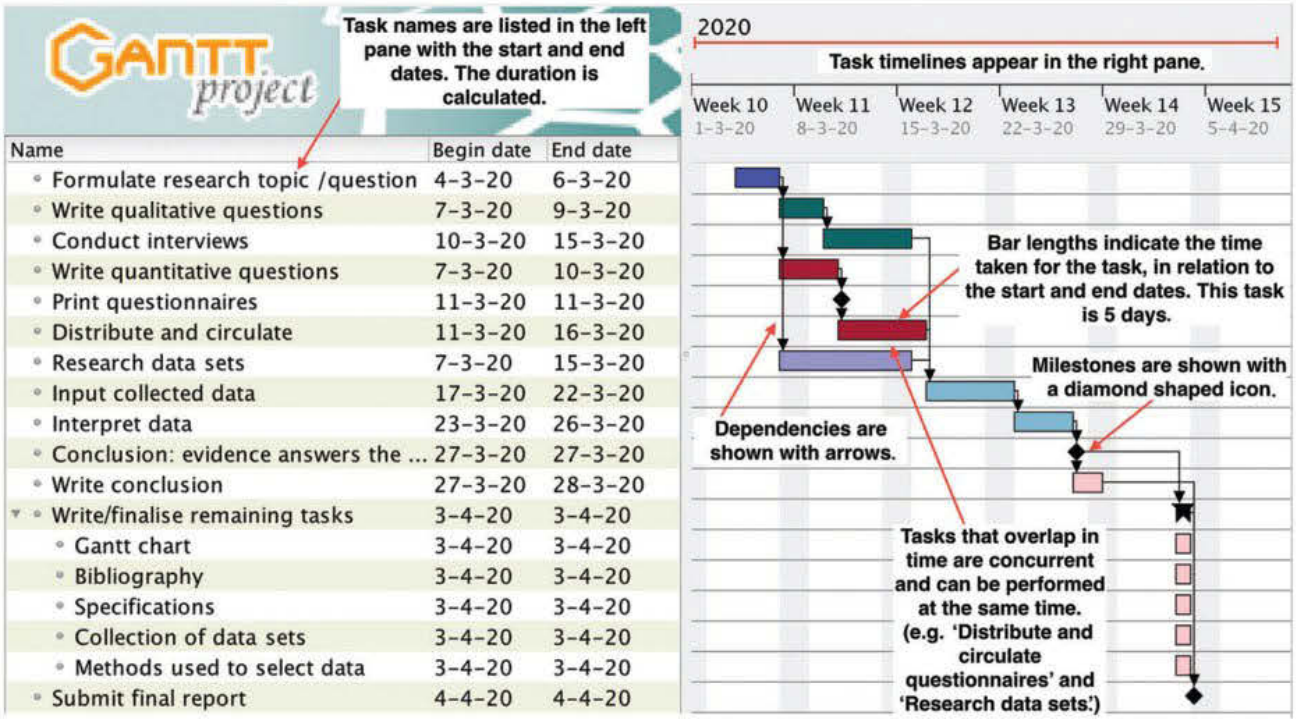
While you can use Gantt chart software to identify the critical path, in the example from Figure 3.4, the critical path comprises the following.

- Formulate research topic/question.
- Write qualitative questions.
- Conduct interviews.
- Input collected data.
- Interpret data.
- Write conclusion.
- Write/finalise remaining tasks.

Sometimes, more than one critical path is possible. No task on the critical path can have its duration changed without affecting the end date of the whole project.

Ad NelsonNet additional resource: Figure 3.4 Sample GanttProject chart, page 121

9780170440875



Name	Begin date	End date
Formulate research topic /question	4-3-20	6-3-20
Write qualitative questions	7-3-20	9-3-20
Conduct interviews	10-3-20	15-3-20
Write quantitative questions	7-3-20	10-3-20
Print questionnaires	11-3-20	11-3-20
Distribute and circulate	11-3-20	16-3-20
Research data sets	7-3-20	15-3-20
Input collected data	17-3-20	22-3-20
Interpret data	23-3-20	26-3-20
Conclusion: evidence answers the ...	27-3-20	27-3-20
Write conclusion	27-3-20	28-3-20
Write/finalise remaining tasks	3-4-20	3-4-20
Gantt chart	3-4-20	3-4-20
Bibliography	3-4-20	3-4-20
Specifications	3-4-20	3-4-20
Collection of data sets	3-4-20	3-4-20
Methods used to select data	3-4-20	3-4-20
Submit final report	4-4-20	4-4-20

FIGURE 3.4 An annotated Gantt chart. This Gantt chart provides a partial sample model for Unit 3, Outcome 2 with placeholder dates.

Documentation using Gantt charts

You can use the Gantt chart you develop to mark your progress throughout your Outcome. By including information about the progress of the task and the planned versus the actual duration of the task, you will help to keep yourself on track.

To manage your solution effectively as a complex project, use your Gantt chart to document the resources you have allocated to it, such as any tools and equipment. While a company or organisation may list consultants and buildings as resources, your resources might be computers, particular data sets, and software tools.

You should also frequently modify your Gantt chart to reflect contingencies. A contingency is an unforeseen event, incident or emergency. You may find that your interview subjects suddenly become unavailable or unwilling to be interviewed. The data set you wish to use is no longer available. The website hosting your questionnaire crashes and loses your data. Your Gantt chart should show problems like this and how you react to them, such as finding new interview subjects, different data sets, or switching to paper-based questionnaires.

You should keep your Gantt chart updated throughout both Unit 3, Outcome 2 *and* Unit 4, Outcome 1. You will submit an initial project plan, indicating times, resources and tasks in Unit 3, Outcome 2. After modifying the plan to indicate changes, you will submit an assessment of the plan in Unit 4, Outcome 1.

THINK ABOUT DATA ANALYTICS 3.1

Project management tools are useful to find the perfect number of people needed on a task so it is finished as quickly as possible without anyone being idle. Use software to develop a Gantt chart to plan the baking of a cake. Assume you can use as many cooks as you want.

THINK ABOUT DATA ANALYTICS 3.2

In terms of project management, research the meaning of:

- an 'optimistic' task duration
- a 'pessimistic' task duration.

SCHOOL-ASSESSED TASK TRACKER

☑ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

CASE STUDY

Gantt chart for creating a database

A small company of web developers are working on a database project together. As part of their project, they need to build a Gantt chart.

Task identification

They first identify the tasks they need to complete using a WBS diagram:

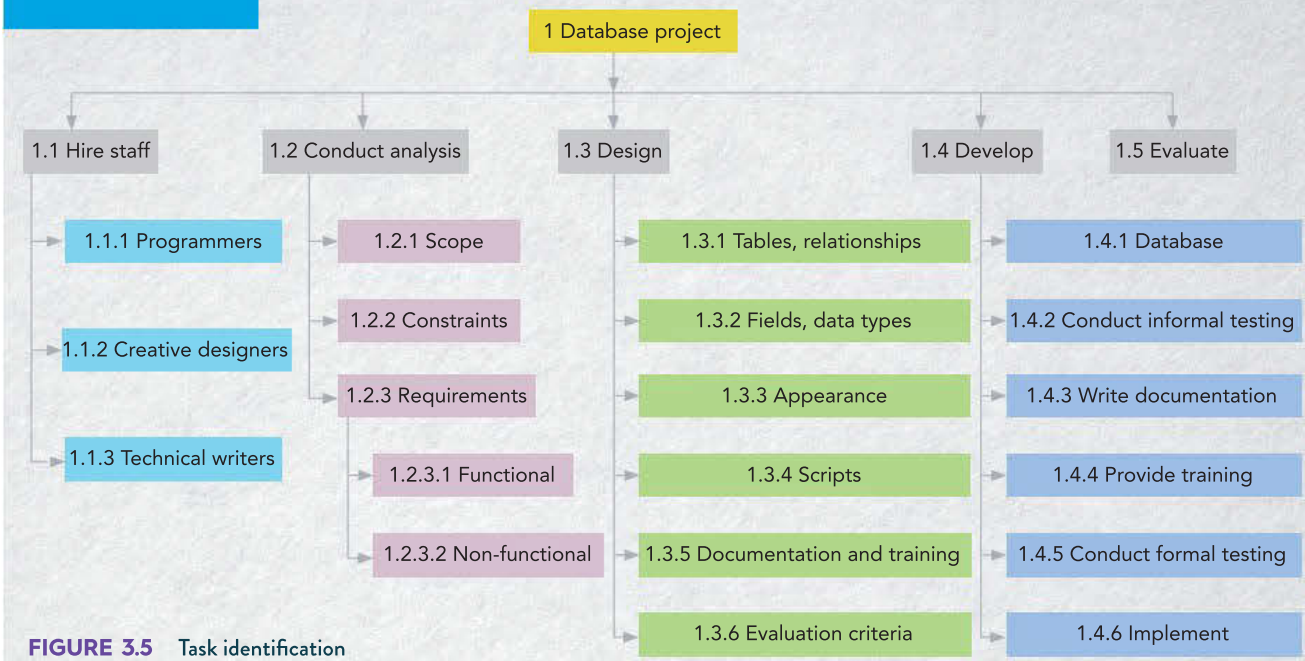


FIGURE 3.5 Task identification

Next, they enter these tasks into their chosen Gantt software, GanttProject:

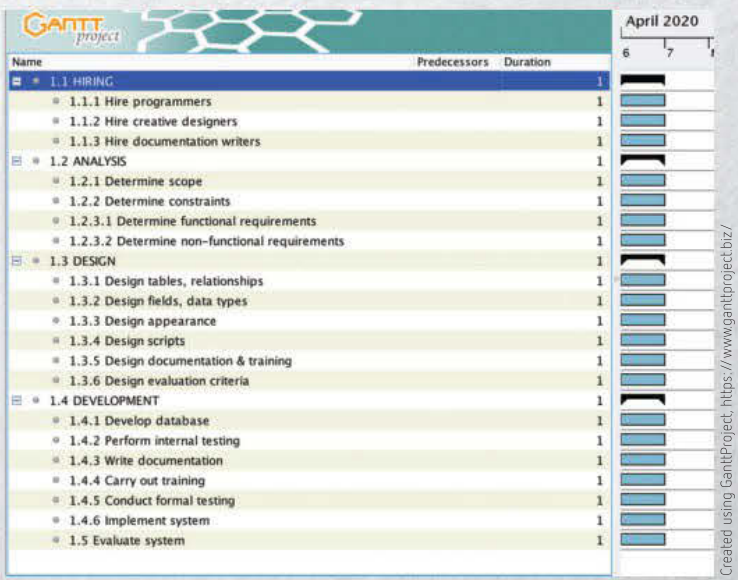


FIGURE 3.6 Entering tasks into Gantt software

They use a hierarchical structure to group tasks under headings, such as HIRING and ANALYSIS to make task management easier. Groups of tasks can be collapsed or expanded or moved as a group. In GanttProject, multiple levels of subtasks can easily be created by just indenting them in the task properties.

9780170440875

Sequencing

The tasks are sufficiently sequenced, but the order can be shifted easily if needed. The developers start creating dependencies, forcing dependent tasks to wait until their predecessors have finished.

- All of the HIRING (1.1) tasks can start immediately.
- Management will complete the tasks in the ANALYSIS group (1.2), which can begin immediately and run concurrently with the hiring tasks.
- The DESIGN (1.3) tasks cannot begin until the ANALYSIS tasks are complete, so DESIGN is made dependent on ANALYSIS.
- The DEVELOPMENT (1.4) tasks cannot begin until DESIGN is finished, so DESIGN is added as a predecessor to DEVELOPMENT.
- 'Evaluate system' must wait for everything else to finish, so it is made dependent on DEVELOPMENT.

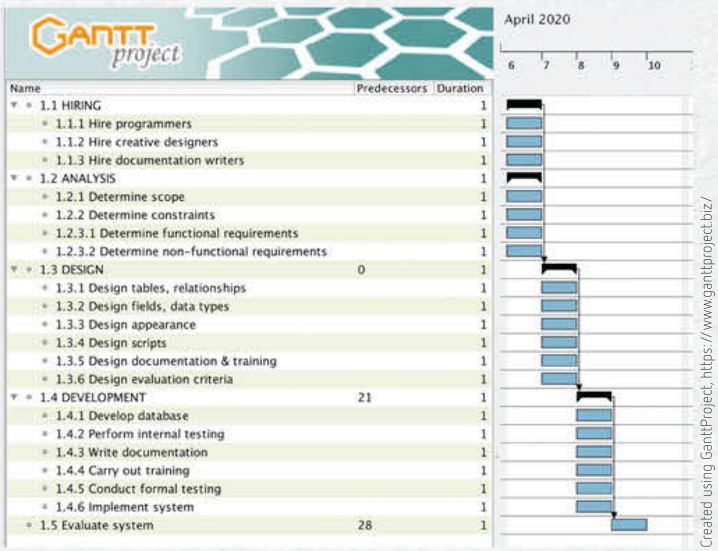


FIGURE 3.7 Major dependencies have been added. Arrows lead from predecessor tasks to dependent tasks.

Subtler dependencies can now be added.

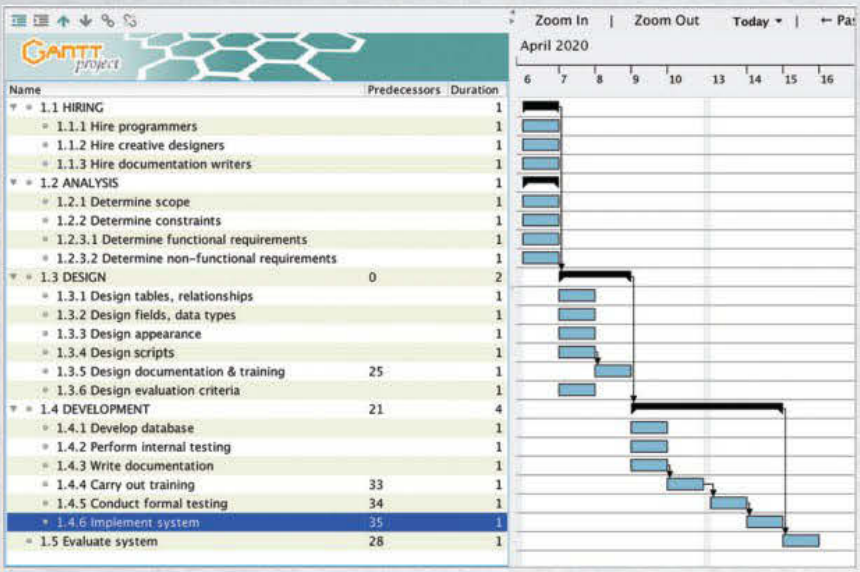
- The team wants DESIGN to be well underway before they start creating documentation and training, because this would be greatly affected by changes to the database. They make 1.3.5 ('Design documentation & training') dependent on other design tasks being finished.
- They also make 1.4.4 ('Carry out training') dependent on 1.4.3 ('Write documentation') being finished.
- 'Conduct formal testing' (1.4.5) must follow all database creation tasks, so they add another dependency.
- 'Implement system' (1.4.6) comes as the last stage of development.

They are now happy with the logical task sequences and dependencies, but they know that if their needs change later, they can still easily adapt the chart to suit their needs.

Time allocation resources

The project developers now tackle the challenge of the time required for each task. They consult with experts and colleagues, and use their extensive experience and knowledge of past projects to guide their estimates.

9780170440875



Ad
NelsonNet additional resource: Figure 3.8 Gantt chart showing task durations.

Ad
NelsonNet additional resource: Figure 3.9 Gantt chart finished first draft.

FIGURE 3.8 Dependencies have been created.

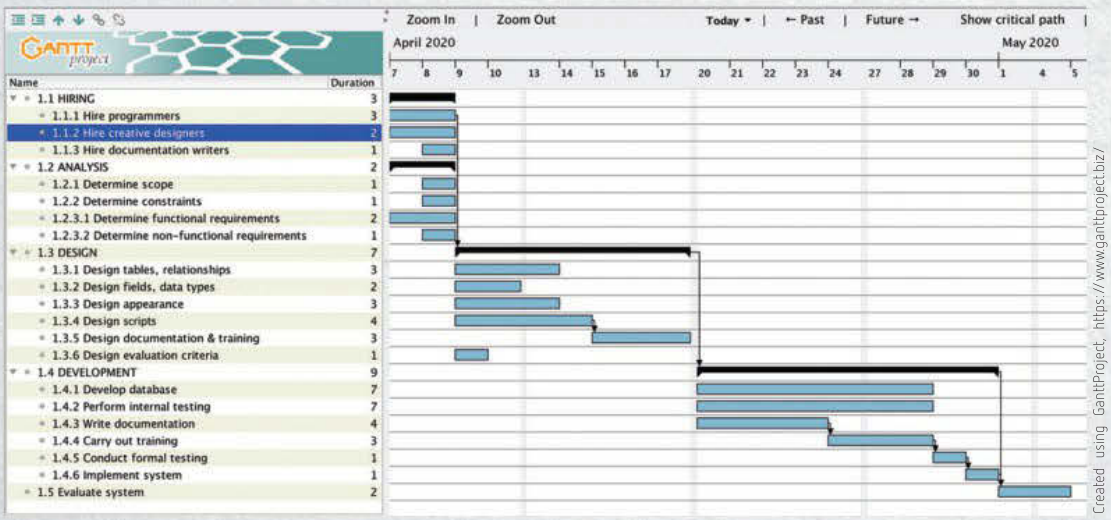


FIGURE 3.9 Task durations have been estimated.

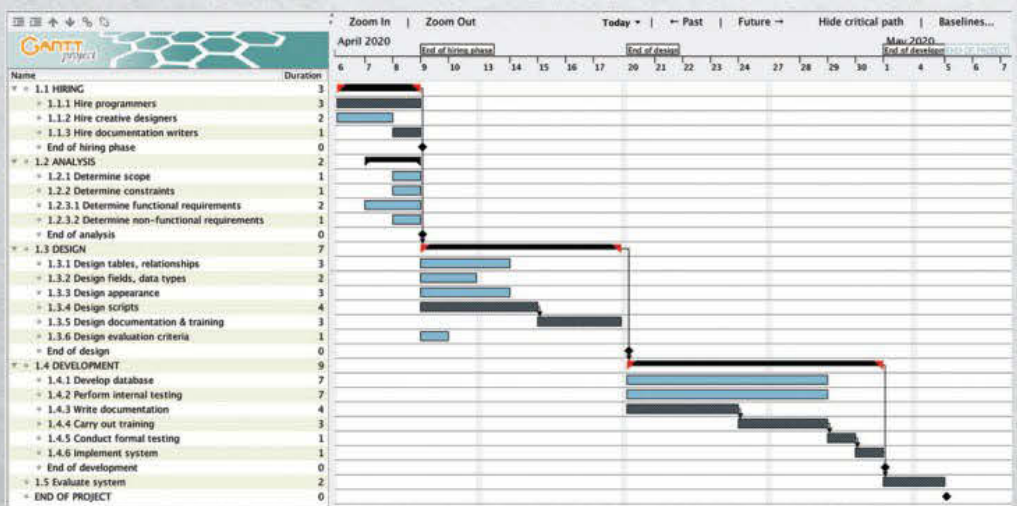


FIGURE 3.10 Finished first draft of the Gantt chart

9780170440875

The developers can finalise their Gantt chart by:

- selecting a project start date
- showing a critical path (in dark grey) by toggling the 'show critical path' button at top right of screen
- adding milestones to mark the end of major stages of the project.

Questions

1 The project team has made a mistake with the starting date of the 'Evaluate system' task. Explain why. How might they fix it?
2 How much slack does the 'Hire creative designers' task have?
3 If designing scripts takes a day longer than expected, would it affect the project end date?
4 The team discovers that the 'Develop database' task is running overtime. How could they keep the project from running past its planned end date?
5 Explain the benefits of using Gantt chart software instead of using a pen and paper or a whiteboard.

Activity

6 Obtain a Gantt software tool and create a simple chart of your own.

Documentation using Gantt charts

To ensure project workers are not booked to be in two places at once, or not booked at all, and that equipment is ready at the right time, the project manager allocates resources to tasks using the Gantt chart (Figure 3.11).

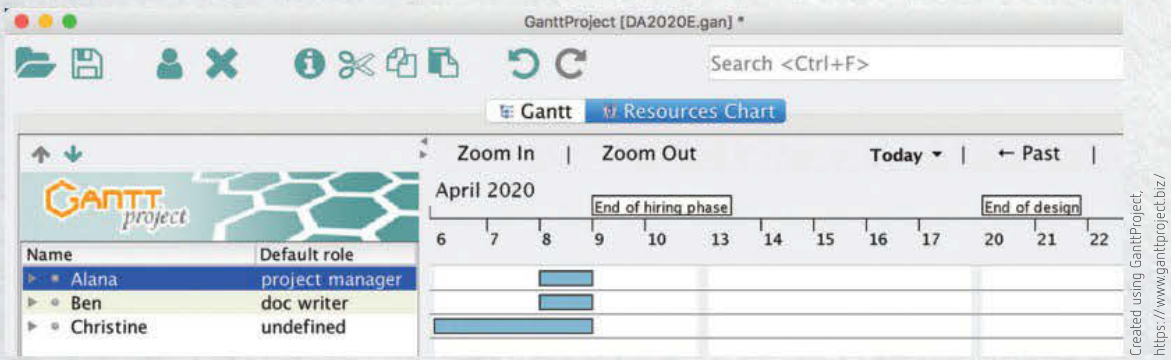


FIGURE 3.11 Adding resources to the project

Once the project is underway, the team will continue to refer to the Gantt chart to monitor their progress, and they will frequently modify and update the chart when contingencies force plans to change.

While Gantt charts are one crucial aspect of project management, good file management practices are another. Wise file-naming strategies are easy to learn and useful in many ways. If you name files well, you will find it easier to keep track of the materials you collect for your solution.

9780170440875

Why must we begin with a research question?

In Unit 3, Outcome 2, you must undertake data collection, in particular selecting, referencing, organising, manipulating and interpreting data to determine the findings for your research question.

The steps for collecting and analysing data can be presented as follows:

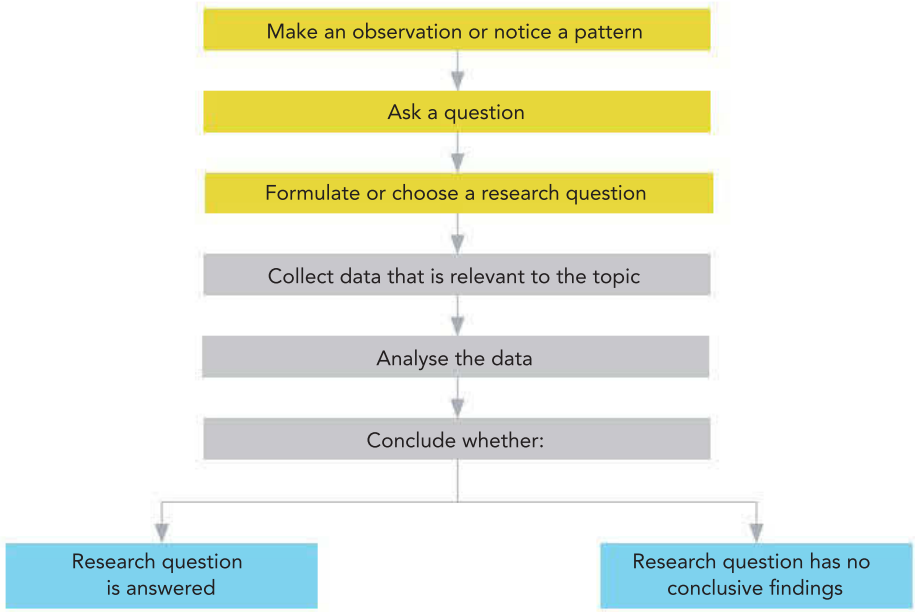


FIGURE 3.12 Researching, analysing and determining research findings

Care must be taken to ensure sufficient data can be collected and analysed, and findings can be stated. A vague research question will not yield findings with insights, or anything not previously known. The research question you formulate for your Outcome will have an impact upon the data you gather and your solution. The nature of the research question you choose will determine the data types you need. It will also have an impact upon the solution you must create for Unit 4, Outcome 1, so make sure you choose a reasonable research question that identifies a specific statement to be the basis for your investigations.

Your research question can be on any topic. However, you will find the SAT much more engaging and interesting if you choose a research question that relates to an interest of yours.

TABLE 3.1 Features of a reasonable research question

Feature	Reason	Examples
Your choice of topic should be able to generate a specific question.	Clarity is needed if you are to convey any findings.	
The specific question should identify variables. An independent variable is the cause of the change, and a dependent variable is the factor that is affected. Researchers change the independent variable to observe the effects on the dependent variable.	Both the independent and dependent variables need to be measurable if the findings are to be valid and presented visually (that is, graphed).	Is attendance at AFL dependent on ticket prices? Lower prices (independent variable) encourages spending (dependent variable) – Data set: AFL archives. What is the safest time to be a car driver, or passenger, in Victoria? Recorded injuries (dependent variable) occur at certain time of day/on certain days (independent variable) – Data set: vic.gov traffic.

9780170440875

Feature	Reason	Examples
It needs at least some supporting evidence or observation.	Unsubstantiated supposition is not a question that can be answered, regardless of the charisma, status or dedication of the believer. A research finding with no evidence is conjecture, speculation or faith.	'Do crystals promote healing?' – Clearly not supported by evidence. 'Do maggots have beneficial medical applications?' – This is supported by current and historical medical evidence. 'Is the Earth flat?' – This is not supported by evidence. 'Is climate change occurring?' – Supported by current and historical evidence. 'Was there a moon landing in July 1969?' – Overwhelming documented evidence is historically available.
Evidence should be objective, not subjective.	Subjective evidence is opinion, not fact.	'Are dogs better pets than cats?' – Opinion. 'Does pet ownership have beneficial medical effects for people?' – Supported by evidence.
It is not trivial.	Insignificant research topics have little merit.	'Are brontosauruses thin at one end, much thicker at the middle and thin again at the far end?' – Is objective, but insignificant. 'Were brontosauruses warm-blooded?' – Would be significant because it would have evolutionary and biological implications.
It should be able to make testable predictions.	A successful prediction provides strong evidence that the dependent variable is in fact affected by the independent variable.	Moore's law states that the number of transistors in an integrated circuit doubles approximately every two years. A testable question – Will the number of transistors in a CPU double within 24 months?
It should be no more complicated than need be to explain an observation.	Researchers use Occam's razor, a principle that states that when there are competing explanations for an observation, the simpler one with fewer assumptions and conditions is preferable.	To explain crop circles – Is it more likely that (a) aliens travelled thousands of years to interfere with our crops and then fly away, or (b) some bored local farmers had some fun?

Formulating your research question

One way of formalising your research statement is as follows:

Choose a topic you find interesting, perhaps from one of those suggested on in Table 3.1. Think about an interesting question that may be true about the topic, such as:

- Do male professional athletes typically earn more than female professional athletes?
- Is the consumption of genetically modified food increasing steadily in Australia?

9780170440875

- Will the desktop PC become obsolete and soon disappear?
- Is Buddhism rapidly growing in popularity in Australia?
- Are more and more people becoming vegan in Australia?

Conduct preliminary research to locate relevant sources of online and printed data on your topic to see if it is worthwhile pursuing for your Outcome. Restructure your topic into a testable, reasonable question for your research.

Alternatively, if you find it difficult to choose a topic, consider looking at interesting data sets online until you find one that you find intriguing.

There are many complex data sets available online. The weblinks suggested here are reputable and have been verified as accurate. Care is necessary to ensure the circumstances of your data meets requirements.

Other topics or data sets could be:

- names and times of all participants in the Melbourne marathon*
- daily electricity consumption recorded by a smart meter*
- traffic volumes at all main intersections in Melbourne
- top 100 music trends
- sporting team success.

*Certain privacy considerations must be followed if you choose this topic. Privacy will be discussed in Chapter 7.

You could also scan news websites until an issue grabs your attention. You could also construct your research question from something that is relevant to you, such as local sporting culture or emerging trends in your community.

Victorian Government data – search for speed camera data or location and time of car 'accidents' in Victoria

City of Melbourne Open Data project:
Parking
City sensors
Environment
Transport

Australian Bureau of Statistics

Bureau of Meteorology

Proof, support and refutation

A research question can sometimes never be proven true with data: it can only ever be supported. However, any statement can be refuted (proven to be false). A reasonable research question is not like a debate topic with equally valid opinions on both sides. You must be able to objectively support it or refute it.

Up until 1697, it was presumed that black swans did not exist, seeing that, broadly speaking, in recorded history every swan anyone had ever seen until that point had been white. However, after Willem de Vlamingh sailed to Australia in 1697, and saw his first black swan, this changed. The presumption that 'all swans are white' was refuted.

A thing can never be proven to be impossible. History is filled with instances of people who declared that many now-commonplace things were impossible:

> 'Rail travel at high speed (of 30 miles per hour) is not possible because passengers, unable to breathe, would die of asphyxia.'
> – Dr Dionysius Lardner, 1828

> 'This "telephone" has too many shortcomings to be seriously considered as a means of communication. The device is inherently of no value to us.'
> – William Orton, President of Western Union, 1876

> 'X-rays will prove to be a hoax.'
> – Lord Kelvin, President of the Royal Society, 1883

> 'A rocket will never be able to leave the Earth's atmosphere.'
> – *The New York Times*, Editorial, 1921

A 'black swan moment' is when an event that is unexpected and without precedent occurs.

9780170440875

'I think there is a world market for maybe five computers.'
– Thomas Watson, President of IBM, 1943

'The world potential market for copying machines is 5000 at most.'
– IBM, to the eventual founders of Xerox, saying the photocopier had no market large enough to justify production, 1959

'We don't like their sound, and guitar music is on the way out.'
– Decca Recording Company on declining to sign the Beatles, 1962

'There is no reason anyone would want a computer in their home.'
– Ken Olsen, founder of Digital Equipment Corporation, 1977

'No-one will need any more than 640 KB of RAM.'
– IBM Engineer, 1992 (sometimes wrongly attributed to Bill Gates, founder of Microsoft)

'There is no chance that the iPhone is going to get any significant market share.'
– Steve Ballmer, Microsoft CEO, *USA Today*, 2007.

When you collect your data, make sure that your research question and data are testing what you think you are testing, and nothing else is influencing your result. To do this, you should try to conduct a **fair test**. To create a fair test, you need to understand variables. A variable is a factor that can change. In a typical experiment or investigation, several variables may affect your results. As you collect your data, you must make sure that you plan a fair test so that only one variable may change and the others are kept the same, or are **controlled**. (Independent and dependent variables are also described in Table 3.1 on page 124–5.)

If you are not careful and do not control your variables, you will not be sure if your data supports your research question or makes a conclusive finding.

CASE STUDY

Pavlov's dogs

One day, Pavlov, a hospital worker, notices patients playing with a dog that a visitor has brought in during visiting hours. When he approaches to reprimand the visitor for bringing an animal into the hospital, he stops himself, because he realises that many of the patients are smiling and seem to be free of their usual pain.

Finding this interesting, Pavlov forms a general question:

Does interaction with animals affect the mental health of hospital patients?

While this is interesting, the question is vague and unfocused. It is not really a research question at all. Pavlov makes it more specific and targeted:

Does interaction with animals improve the mental health of patients?

Now there is some direction, but the research question is not directly testable and verifiable. Pavlov is curious whether he can verify his hunch. He wonders why animals improve the mental health of patients, and how his research question can be supported. He devises a testable question that may lead to research:

Does interaction with animals improve the mental health of patients because it reduces their stress levels?

9780170440875

This has now established variables that can be modified and observed.

- Independent variable: Interaction with animals can be modified by controlling the amount of interaction patients have with animals.
- Dependent variable: Stress levels can be measured by conducting interviews or measuring stress hormones in patients' blood.

Pavlov realises that when he collects data, it is important that other variables that could affect the results are controlled. What if, rather than the pets affecting stress levels, other factors were responsible, such as illness, diet, sex, location, age and so on? Data should be filtered so that only comparable data is actually compared. Pavlov realises that it would be futile to compare the data from 90-year-old dementia patients in Germany *with* pets and with children with cancer in China *without* pets. Too many uncontrolled variables such as age, location, diet, and so on, could affect the data.

Once Pavlov stated his research question clearly, he could generate a prediction from it:

> A patient's stress levels will be reduced if the patient is given exposure to domesticated animals, such as cats or dogs.

With interest from hospital administrators, experiments were conducted, primary data was collected from interviews, and blood tests and secondary evidence were collected from similar research elsewhere. The data was averaged and compared. A statistical significance test led Pavlov to the conclusion that, on average, there was a statistically significant improvement in the mental health of patients who were given regular exposure to animals compared to those without exposure.

Pavlov's research question was supported, and the hospital administration used the information to create and fund a 'Pets As Therapy' program.

Interpreting your data

Interpreting your data may be a complex process depending on the question you have chosen. You may find it difficult to be sure whether your question has been supported or refuted.

When you are interpreting your data, consider the following questions.

- Do your results make sense?
- Could your results be interpreted differently?
- Is your supporting data of good quality?
- Is your supporting data current?

You could also consider using tools for organisation and techniques for finding patterns, such as **mind mapping**. Patterns in data can be identified more easily if you use tools such as **graphs** or other visual techniques such as **frequency distribution tables**, which will show the number of observations in each category.

Primary and secondary data

As you learnt in Chapter 1 (page 5), data that has not been filtered by interpretation or evaluation is called **primary data**.

Secondary data differs from primary data because it has been interpreted or manipulated by someone *other* than the researcher.

When collecting data, the origin of the interpretation tells you whether it is primary or secondary, but this is not the only type of categorisation that applies to data. Data is also categorised in other important ways, such as whether the data is quantitative or qualitative.

9780170440875

RESEARCH

Potential research question and data sources

If you have not decided on a research question yet, use this activity to help you.

1. List three general topics that you think you may be interested in using for your focus in the table. You will only end up using one of them, but it is good idea to have a few in mind just in case one or more are not suitable.
2. Brainstorm the potential primary and secondary data types for each of your topics and set them out in a table like that presented here. They do not need to be correct or exhaustive. You are just trying to imagine the possible types of data you could gather for the topic if you end up choosing to formulate a question around it, such as interviews, photos, journals, newspaper articles, interviews and so on.

Potential research topic:		Potential research topic:		Potential research topic:	
Primary data sources	Secondary data sources	Primary data sources	Secondary data sources	Primary data sources	Secondary data sources

Quantitative and qualitative data

When you collect primary and secondary data to support your research topic or question, it will be made up of quantitative or qualitative data, or perhaps both. Both quantitative and qualitative data serve a purpose. It will be up to you to choose the appropriate type of data for each occasion.

If you do not know much about a topic, but you want it to be the subject of your Outcome, you may consider starting with some small-scale qualitative research. It will help you to hone in on what interests you and what other people generally think and feel about that subject, and it will give you ideas for the quantitative research you need to conduct later.

The simplest way to distinguish between quantitative data and qualitative data is to think of them as *quantity* and *quality*.

Quantitative data is concerned with numbers, measurement and being scientific by taking observations. It uses an objective approach, **closed questions**, and can be easily measured and scored. This could include temperature records for Melbourne during summer, maths test results, and attendance numbers for AFLW matches.

Qualitative data is expressed in words because it is concerned with feelings, personal views and experiences, and opinions. Qualitative data is more difficult to analyse statistically because it has to be coded to be scored. It is subjective.

Consider the following research question: *'Do increasing numbers of people in Melbourne eat more vegetables because they believe it has health benefits?'* You would need to collect both quantitative and qualitative data for this research question: quantitative data to confirm whether increasing numbers of people are, in fact, eating more vegetables in Melbourne, and qualitative data to determine if it is because they believe it has health benefits.

THINK ABOUT DATA ANALYTICS 3.3

Categorise the scoring of the following Olympic events as either qualitative or quantitative, and explain your reasoning for each choice.

1. Figure skating
2. Speed skating
3. 100 m sprint
4. Judo
5. Team show jumping (equestrian)

SCHOOL-ASSESSED TASK TRACKER

☑ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

Quantitative data	Qualitative data
Expressed numerically	Expressed in words
Acquired by measurement	Concerned with opinions, feelings, motives and preferences
Uses scientific approach	Holistic (total picture) approach
Can be easily analysed statistically	Needs to be encoded to be analysed statistically
Collectable in large quantities	Collected on a smaller scale
Objective facts, such as, '8% of RubyMede students play netball for at least two hours per week'	Subjective facts, such as 'I play netball for two hours each week because it helps me to socialise and de-stress, and I don't have to think about schoolwork or or exams'
Can be compared with historical data	Can inform policy development
Can be used to check current status of an ongoing issue	Can be used to delve into ethical or moral issues

FIGURE 3.13 The differences between quantitative and qualitative data

THINK ABOUT DATA ANALYTICS 3.4

After reading the dieting section of a popular magazine, you ask the question 'Do people only like to eat "superfoods", such as beetroot, because they are fashionable and that no-one really likes to eat them?'

Describe the sorts of quantitative and qualitative data you would need to find to generate data to answer your question.

The following section discusses coding qualitative data. You will find this useful for your Outcome because the issue you discuss and choose to frame your research question around may relate to emerging trends and shifting patterns. Trends and shifting patterns reflect opinions, which are subjective (qualitative) in nature.

Coding qualitative data

Quantitative data is collected using techniques such as online questionnaires or surveys that feature **Likert scales** (for example, 1–5 or 1–10 scale), multiple-choice dropdowns and/or radio buttons. It takes a highly structured, digital and numeric form. This makes it easy to score and process, and thus convert into **information**. It can be a simple process to unlock the potential value of quantitative data.

A **Likert scale** (pronounced lik-ert rather than ly-kart) is a psychological measurement tool that is used to gauge attitudes, values and opinions. It functions by having a person complete a questionnaire that requires them to estimate the extent to which they agree or disagree with a series of statements. The Likert scale is named after its creator, Rensis Likert, who developed it in 1932.

The same cannot be said for qualitative data. Even if qualitative data is gathered online, when you answer a qualitative question, it will often be the type of question that gives the respondent a textbox to write their answer. Qualitative data is non-numerical. Answers are expected to be freeform. The idea is that when you ask qualitative questions, you want respondents to tell the *complete* truth, so you do not limit the length of answers or the form, patterns or vocabulary they will take. The result of this is that you cannot score it easily.

After collecting qualitative data for your Outcome, you will most likely need to digitise it and 'code' it so you can process and use it as information.

How do you turn this data with all of its potential value into useful information when it cannot be scored very easily? The answer is that you must transform the data by interpreting and coding it into a summarised form that will help you to analyse it appropriately. The summarised form could be a label, category or simply a number.

Do not try to invent rules for how each and every answer can be transformed into a simple code. This may be unproductive, and would be a very time-consuming task, because qualitative answers truly can be massively unpredictable in the forms they may take. You might also unintentionally introduce bias and error that may skew the information.

9780170440875

Applying categories or labels may be more helpful than applying numbers to certain types of data. It is important to consider the data output and also the question asked when deciding how to process the data.

Qualitative coding can be valuable in creating categories. Until the survey is conducted, researchers may have little idea of the general categories of problems. An example is provided in the following case study.

Quantitative research can then apply predetermined categories. It usually does not allow respondents to add new categories.

Coding can reduce the original wordiness to a more manageable form using freely chosen summary terms, and this is sometimes called **descriptive coding**.

Coding can also replace entire ideas with symbols from a fixed set of summary numbers, codes or labels, and this is known as **analytic coding** or **theoretical coding**.

Encoded qualitative data allows responses to be grouped, averaged, totalled and compared. Once labels and categories have been assigned, it is easy to find individuals who meet certain criteria, and to determine patterns.

Qualitative data coding

RESEARCH

Consider this question asked of politicians:

> What are your views on human-caused (anthropomorphic) climate change?

1 Each response to the question above is listed in the table below. Code each response in the right-hand column as a number from 0 to 5, with 0 = strongly support anthropomorphic climate change to 5 = strongly object to the idea that humans could have caused climate change.

Response	Coded as
The climate has always been changing and recent events are just weather, not climate change.	
Global warming is caused by the Sun, and all the planets are warming.	
Without human activities, the influence of natural factors alone would actually have had a slight cooling effect on global climate over the last 50 years.	
The climate models are unreliable; 1932 was the hottest year and it's been cooling ever since.	
We're heading into another ice age.	
The science is settled.	
I don't know, it's all too confusing.	
Greenhouse gas emissions from human activities are the only factors that can account for the observed warming over the last century.	

Source: URL https://www.skepticalscience.com/argument.php. Accessed December 2018.

2 What criteria did you use to turn a thought into a single number? Did you find it difficult to apply the numbers?

3 The detail and nuance in the responses are stripped back by coding each response as a number. There are other ways of coding qualitative data, such as categorising and labelling. How might you have coded the responses to better represent their content?

9780170440875

CASE STUDY

Ready Set Go! gym

A group of 28 people who all belong to the Ready Set Go! gym indicate to management that they do not want to renew their membership within the same month. The manager of the gym, Kym Carr, is alarmed by this news.

While it is normal policy to try to upsell a new membership, or send out a (quantitative) questionnaire, this has not been very successful. The upselling is typically met with aggression or avoidance and the questionnaires do not get returned. Kym has been taking some psychology courses that have taught her a bit about statistics and interview techniques, so decides to do something different – she schedules an interview with each member, in which she collects qualitative data to find out why they are leaving the gym. When scheduling the interview, she indicates that she wants to fix the problems with the gym and will not try to sell them anything unless it seems like there is something feasible that they may want.

She asks each member the following two questions only, and lets each member talk as much as they want.

1 Why are you leaving the gym?
2 How could we convince you to stay?

Each interview lasts for approximately half an hour and, to Kym, it feels very private and different to the appointments with members that she usually has. She feels as though the members are opening up to her in ways they normally would not. At certain points, she feels almost as if she is hearing their secrets. She takes notes, but does not interrupt them unless she needs them to further explain something she feels she may have misunderstood.

> When I first started I had a lot of enthusiasm and energy, you know, but as the weeks went on, I was coming by myself and it just got harder. And then I weighed myself and nothing was really happening, and that made it worse. I feel pretty isolated when I do come in, and I'm sorry, but you've got that one really rude guy who works here and makes fun of clients within earshot of other people. I find myself worried I'm going to come in on a night when he's working and he's going to see me and make fun. I mean, you realise that's just not on, right? The only machines I really like are the cross-trainers and the rowing and they're not looked after properly. I just feel a bit let down. I feel like crap when I come here so I don't want to come here. I'm meant to come here so I can start to feel better about myself, not worse. And it's not like I have anyone to talk to about it.

After all of the interviews, Kym makes a spreadsheet list of the basic reasons why the members want to leave the gym.

She notes that during the interviews, nine of the people who wanted to cancel their memberships had expressed deep disappointment over the employee who had been unprofessional and linked their lack of results to him. One of the members who mentioned depression also mentioned lack of results, cost of membership and unprofessional behaviour of gym staff.

Kym decides to apply categories to the data to make it easier to understand the issues.

9780170440875

	A	B	C
1	Reason	Mentioned	Notes
2	The cost of membership	5	Maybe do 20% discount as incentive?
3	Poor state of equipment	10	Which equipment? X-trainers, bikes in spin-cycle rooms, some of the barbells, rowing machine, stepper
4	Unpleasant smell from weight room	13	Apparently it smells a bit like mould but it's pretty potent
5	Too much chlorine in the spa and pool	4	Member actually said it hurts his eyes
6	Lack of appealing classes	15	Shortage of good yoga, pilates etc... Not enough variety in the aerobics classes and also we just don't do enough in the aquatics classes full stop
7	Unprofessional behaviour from staff	12	Eric is mentioned 9 times – he needs to go. Notify HR.
8	Low motivation	7	
9	Depression	2	
10	Lack of results	15	Not enough increased muscle mass or significant weight loss within expected timeframe. Why aren't my staff monitoring this? Link to depression and low motivation maybe?

FIGURE 3.14 Reasons for leaving the gym

13	Reason	Mentioned	Categories
14	The cost of membership	5	economic concerns
15	Poor state of equipment	10	physical environment; safety (high priority)
16	Unpleasant smell from weight room	13	physical environment
17	Too much chlorine in the spa and pool	4	physical environment; safety (high priority)
18	Lack of appealing classes	15	physical environment; lack of support; lack of variety
19	Unprofessional behaviour from staff	12	staffing concerns; safety (high priority)
20	Low motivation	7	personal concerns
21	Depression	2	personal concerns
22	Lack of results	15	personal concerns

FIGURE 3.15 Reasons for leaving (with categories applied)

25	Reason to stay	Mentioned	Categories
26	Free membership	9	economic concerns
27	Fire Eric	9	staffing concerns; safety (high priority)
28	Fix the equipment	10	physical environment; safety (high priority)
29	Fix the pool and spa	3	physical environment; safety (high priority)
30	Get rid of the smell	12	physical environment
31	More/better classes	12	variety
32	Provide more support	12	physical environment; support
33	Nothing you can do?	15	personal concerns

FIGURE 3.16 Reasons to stay at the gym (with categories applied)

Kym soon realises that the 'physical environment' and safety of the gym was the most significant problem. As a manager in a health-related industry, Kym considers safety to be its own category and automatically views anything mentioned within it as high priority. Applying categories makes a huge difference.

Kym developed some pie charts based on her data as she investigated. The pie chart on the left in Figure 3.17 (page 134) shows safety as its own category. It is not truly correct, because the staffing issues are all a safety issue, and most of the physical environment issues also overlap with safety. The pie chart on the right does not show safety at all. Rather, it groups items by category and allows Kym to assign safety as a priority to them as needed. How else could you graph this data to show safety without providing a misleading tally?

9780170440875

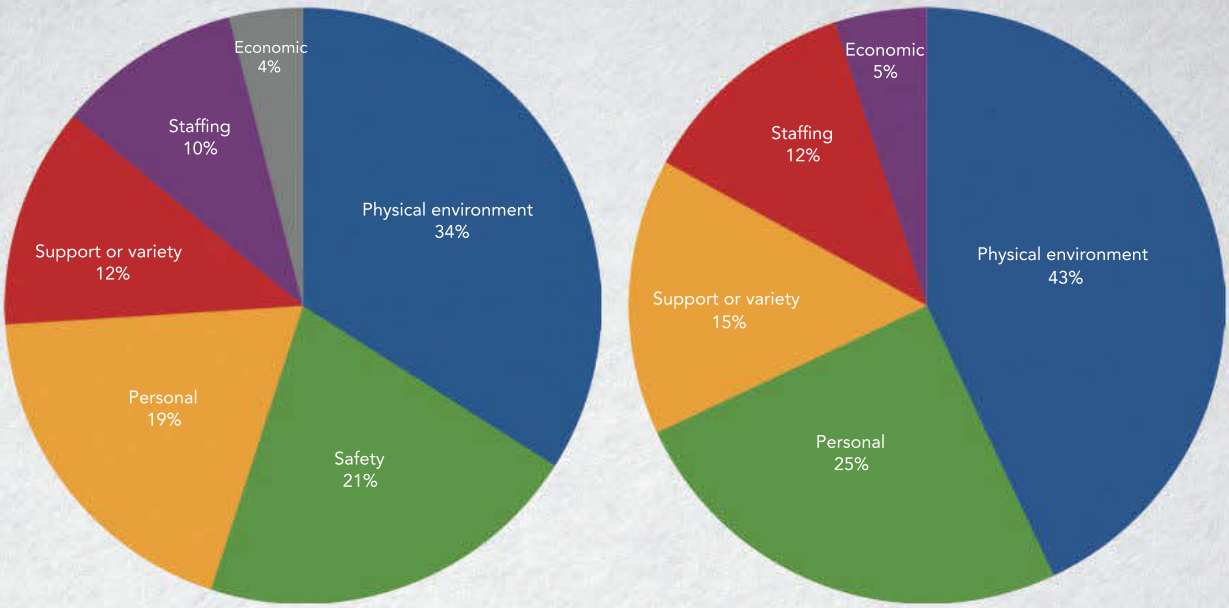


FIGURE 3.17 Kym's data as pie charts

Kym is able to bring in maintenance staff to look at the equipment, investigate the smell in the weight room and deal with the levels of chlorine in the spa and pool.

Another significant category is 'staffing concerns', which also overlaps with safety. Kym decides to remove the employee who has been mentioned repeatedly as problematic because the complaints are very serious and he has been warned twice in the past. She offers a free renewal of memberships to those members who complained about that staff member.

Kym feels a bit despondent that she has so many people falling under the 'nothing you can do' category – personal concerns. She notes that all of the people who said they had low motivation also said they came to the gym by themselves to work out.

After some brainstorming with her staff, Kym contacts all the members with 'personal concerns' (depression, low motivation and lack of results) and asks them if they would consider opting in to her new buddy system program, which pairs members with similar schedules together and also assigns them to a gym instructor as a mentor to monitor their progress and provide them with a more hands-on approach so they do not feel alone.

She also contacts a few local TAFEs and vocational schools to find recently qualified yoga and Pilates instructors to add new classes to the gym, and volunteers to run two more aquatics classes herself each week.

In all, she is only able to convince 17 of the 28 members to keep their memberships, but this is an improvement over the usual percentage, and it has allowed her to weed out an employee who made members uncomfortable, fix broken equipment and help struggling members, which would not have been possible with a standard checkbox questionnaire in the mail.

9780170440875

Examples of analytical coding

1 To find out whether student trust is related to the teacher's experience (Figure 3.18), run a query to find all respondents with the 'Trust' code and get an average of how long their teachers have been teaching. Look up counts of whether individual teachers have less than five years of classroom experience, more than five years of classroom experience, and more than 10 years of classroom experience, to see how often 'Experience' and 'Trust' are commonly combined.

```
1   She cares about my education, so I trust her.
2   Plus, she has loads of experience - she's been around here for
    years and years. She totally knows what she is doing when it comes
    to helping me get good grades on my Outcomes.
3   She always makes time to answer my questions when I get in a
    panic, which is often!
    1 = Trust          2 = Experience          3 = Availability
```

FIGURE 3.18 Descriptive coding of qualitative data about teachers

2 On a scale of 1 to 5, where 5 is feeling good and 1 is feeling not so good, how are you feeling today?

Prompt 1: Statement where respondent has an opinion	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
Prompt 2: Statement predicts behaviour by respondent	Very Likely	Likely	Neutral	Not Likely	Very Unlikely
Prompt 3: Respondent asked to rate their satisfaction	Very Happy	Somewhat Happy	Neutral	Not Very Happy	Not Happy at All
Prompt 4: Respondent asked to rate importance	Very Important	Important	Moderately Important	Slightly Important	Not Important
Prompt 5: Respondent asked – how often?	Almost Always	Sometimes	Every Once in a While	Rarely	Never

FIGURE 3.19 Examples of five-point Likert scales

Common issues with Likert items and scales

The Likert scale data is interpreted once the responses have been assigned a number. Those numbers can be added, averaged and plotted to convey summarised findings that are indicative of the survey responses. Once a chart has been chosen, a visual insight indicates the respondent sentiments on any item.

THINK ABOUT DATA ANALYTICS 3.5

Qualitative coding is difficult, but so is qualitative scoring.

In an examination consisting of multiple-choice and short-answer questions, how is each type of question marked? How do examiners ensure fair marking of short-answer questions? Consider an English examination's essay on a novel: how can the many qualities of an essay be converted into a single number that can fairly be compared with thousands of essays by other students? How can one essay be detected as excellent, competent or unsatisfactory?

SCHOOL-ASSESSED TASK TRACKER

☑ Project plan | ☐ Collect complex data sets | ☐ Analysis | ☐ Folio of alternative designs | ☐ Infographic or dynamic data visualisations | ☐ Evaluation and assessment | ☐ Finalise report or visual plan

9780170440875

The way the questions are presented can sometimes determine or influence respondents in the way the questions are answered. These **skewed results** may come about from **interviewer bias** or **survey distortions**.

Rubric

While a single movie reviewer will have personal, internal criteria for assigning stars, teams of reviewers need to use a shared assessment scheme so their ratings are consistent and comparable. When coding is done by many people, the coded results could be unreliable if they have differing interpretations of answers. Groups of coders, such as VCE exam markers, need to have a clear and shared understanding of how to interpret, evaluate and encode responses. One way to give meaning to codes is to use a rubric: a detailed list of descriptive grading criteria that correspond with a code (the mark). Teachers assess VCE Outcomes in this way. Table 3.2 is an example of a rubric.

TABLE 3.2 Grading criteria rubric

33–40 marks	Thorough and insightful analysis correctly identifies all the website requirements of an online community and acknowledges all relevant technical and nontechnical constraints. All selected design tools are appropriate. Correct design techniques are consistently applied to fully represent the functionality and appearance of a website that is feasible and accordant with the analysis. All manual and electronic validation techniques effectively check the reasonableness of data.
Descriptions are given for 25–32 marks, 17–24 marks, 9–16 marks and finally ...	
1–8 marks	Identifies few website requirements of an online community. Few technical or non-technical constraints are stated. Limited relevant design tools are selected and limited design techniques are applied to outline minimal functionality and appearance features of the website. Few links exist between the design and the analysis. Limited application of manual or electronic validation techniques affects the quality of critical data, and hence, the solution.

Data types and data structures

The data types and structures are discussed in Chapter 1. You will recall that data is categorised into types so it can be stored efficiently and processed effectively. These include the following: numeric (date/time, floating point and integer), text (string), Boolean and character. Basic data types can also be subdivided into more specific variants. Some programs may understand only the 'number' data type, but a programming language may understand a dozen number types.

The basic data types are described in detail in Chapter 1, pages 16–18.

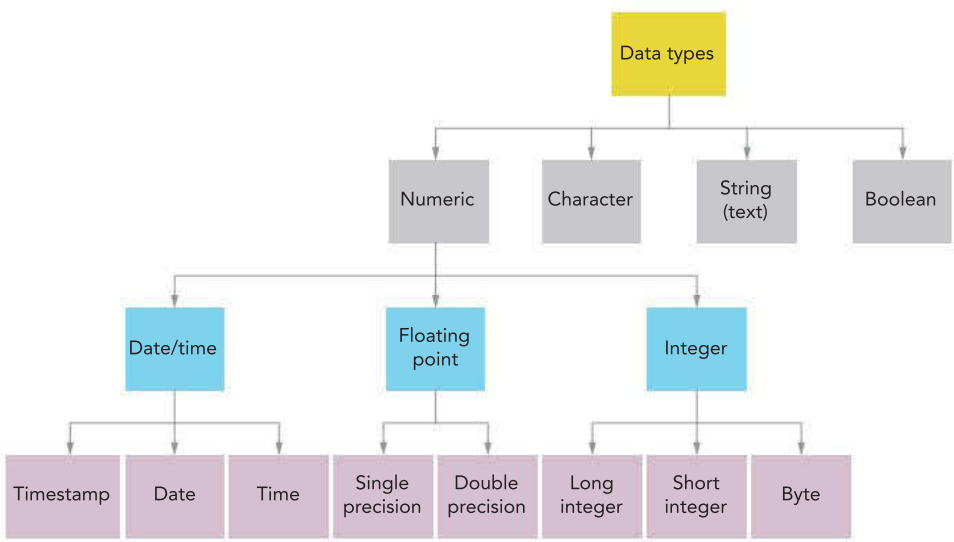


FIGURE 3.20 Common data types used in software tools

9780170440875

Data structures

Databases such as Microsoft Access, LibreOffice Base and FileMaker have formal fields, records and table structures, while spreadsheets do not. In spreadsheets, the nature of the data can remain ambiguous until a specific type of arithmetic or function is used to begin manipulation, such as date, number or text. The cell data type can be specified by several methods. The 'Format' menu can set the data type. Data validation can require the cell to hold certain data types, within a specified range of values. Data structures are important for when charts are being created to represent findings for the SAT. Only when datatypes behave as expected will the chart created display correctly, without inventing odd discrepancies caused by unexpected formatting. For example, if a cell contains text rather than number, a chart will take the cell value as zero. This will affect the presentation of the chart and introduce data that is 'not really there'.

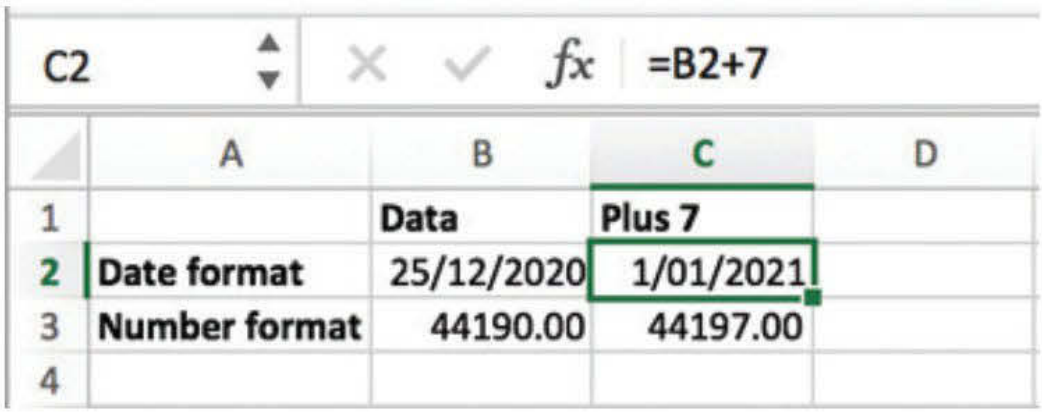


FIGURE 3.21 A spreadsheet handling the same data (column B) as dates and numbers

Referencing data sources

This major idea was introduced in Chapter 1 and you are advised to reread that section since it is relevant to Unit 3, Outcome 4 and Unit 4, Outcome 1. Some of the material is also included here to emphasise the importance of citing your sources correctly.

When any source of information is used in your work, it must be acknowledged. This means not just direct quotes – it also includes ideas, summarising and **paraphrasing**. The sources of intellectual property (IP) that influence your work should be acknowledged. There are several referencing styles, but only the American Psychological Association (APA) style is required for the Data Analytics SAT.

Other referencing styles include the following.

- Harvard
- Chicago
- Institute of Electrical and Electronics Engineers (IEEE)

You are strongly advised to record your sources as you find them. At the end of the Outcome, it will be difficult, and time-consuming, to retrace your steps to locate all of them again to cite them accurately.

9780170440875

Styles

APA style guide

The APA style is expected to be used in referencing for the Outcome. **Citations** appear within the text and their corresponding source details in a reference list at the end of the work. When using the APA style, the author's surname and the publication date are featured in the text.

The APA style treats a website as though the author is clearly distinguishable from any sponsoring website. If a website has no information about the author, you should check the legal section of the website, web page or document.

TABLE 3.3 A simple website reference list citation using APA style

APA	Author or organisation who created the web page (if clearly identifiable).	(Year the web page was published, or most recently updated).	*Title (of the web page or document, in italics if it stands alone)*	[Format description if applicable].	Retrieved from URL.
	Victorian Government – Police department.	(2020).	*Traffic speed camera data*	[Data file].	www.data.vic.gov.au.
	City of Melbourne.	(2020).	*City of Melbourne open data project – parking*	[Data file].	https://data.melbourne.vic.gov.au.
	Australian Bureau of Statistics.	(2020).	*Australian Bureau of Statistics – Census*		www.abs.gov.au.
	Bureau of Meteorology.	(2020).	*Victorian rainfall records*	[Data file].	www.bom.gov.au.

Refer to Chapter 1, page 14 for an example on citing a book in text and Table 1.1, page 15 for the reference list entry.

So, a reference line to the Australian Bureau of Statistics website might look like:

> Australian Bureau of Statistics. (2020). *Australian Bureau of Statistics – Census.* www.abs.gov.au.

Referencing tips

These are repeated from Chapter 1 (page 15).

Quotations

The Citation Machine is an online citation generator that can produce an APA style reference.

- Do not correct errors in quotations. A quotation should be represented as it appears originally, even if it has errors. You can use '[*sic*]' to acknowledge that you have noticed an error in the quotation and that it has not come from you.
- Do not change the spelling in a quotation or a name to match Australian English. Preserve correct names and original spelling. For instance, if you quote the World Health Organization, or need to make note of the World Trade Center, you would not change their names to World Health Organisation and World Trade Centre.
- Use square brackets to add explanatory information or give context to a quotation; for example, Darcy's claim that 'during the [American Civil] war, infections killed more men than bullets ever did …'
- Use ellipses '…' to indicate when part of a quote has been omitted.

9780170440875

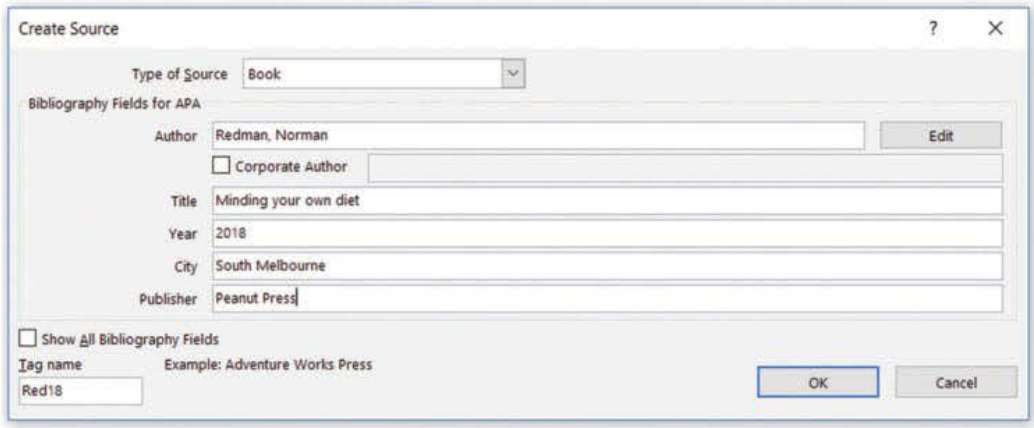


FIGURE 3.22 Example of Microsoft Word references (References>Insert Citation dialog box)

Reference list

- You can use Microsoft Word to save a list of your sources in a document and produce a reference list or bibliography. On the 'References' ribbon, look for 'Bibliography' or 'Citations & Bibliography' (though this will depend on your version of Microsoft Word).
- Write 'n.d.' if no date is given for a website or document.
- If no page number is known, write 'n.p.' and provide an estimated page, paragraph number, or nearby heading.
- Use unspaced en dashes rather than hyphens in your page reference spans. To insert an en dash in Microsoft Word in Windows, hit CTRL and minus on the number pad, or ALT and 0150.
- References are presented alphabetically using the surname of the first author.
- In the reference list, use the hanging indent paragraph style.
- If the source has a **digital object identifier (DOI)**, place it at the end of the reference.

Data integrity

You learnt about the factors influencing the integrity of data in Chapter 1. In this chapter, you will learn about the criteria that is used to check data integrity.

Authenticity

Authenticity relates to how genuine the data set is (Figure 3.23).

The challenge of identifying inauthentic data is becoming more difficult every day. Spoofing attacks occur frequently online. In a spoofing attack, one person (or multiple people) masquerades as another for the purposes of fraud, advantage or amusement. This usually requires falsifying data, such as caller IDs, email addresses and IP addresses.

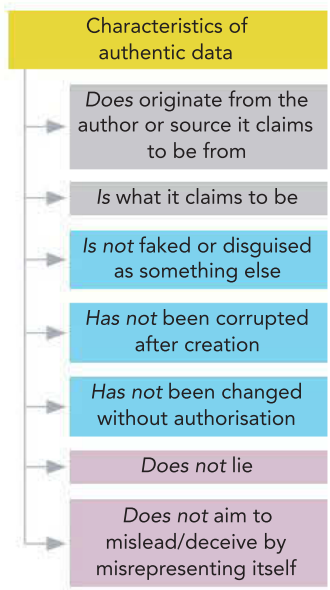


FIGURE 3.23 The characteristics of authentic data

9780170440875

Authors plagiarise other people's publications and steal data. Similarly, copyright, source and author information can be edited out of pictures and documents, making their true source, age and legal status unknown. Watermarks can be removed from digital images.

Legitimate media and advertising productions add to the challenges in several ways. Music producers use auto-tune to repair flat notes their singers have hit during recording. Computer-generated imagery (CGI) in movies and TV is becoming undetectable, while magazines edit photos of models into impossibly perfect specimens. Viral ad campaigns and fake YouTube videos grab attention by appearing to be genuine.

Few people use their real names online. When only avatars and nicknames appear, true identities are uncertain. Authentic social networking accounts get hacked and post fake updates or tweets. Fake celebrity accounts are easy to create. Even the 'From' addresses in email are easy to fake. Images that may have already been clearly identified as fake may be re-posted without the warning.

Torrent sites increase the likelihood of downloading data that differs from the original. Although uploaders who post files that have been cracked and corrupted or changed in some way are unlikely to go unnoticed for very long, in the short term, inexperienced users are vulnerable to downloading the corrupted data. While torrent sites often have legal downloadable files, many include cracked software. Cracked software in itself is a challenge to authentic data, but it also poses both legal and ethical implications for the user. Many users who choose to download cracked software knowingly bypass alerts from their firewall or malware detector to install the program, essentially rolling the dice on the possibility that it may have been changed in other ways.

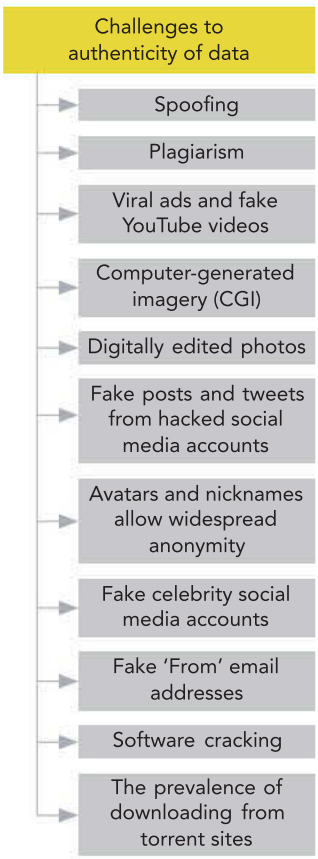


FIGURE 3.24 Challenges to authenticity of data

Authenticity techniques

There are a variety of ways of enhancing the authenticity of data. The technique used will depend on the circumstance. Broadly, the techniques can be categorised as digital or non-digital.

Digital

- Use digital signatures. A digital signature is a mathematical method used to validate the authenticity and integrity of messages, software and digital documents.
- Use Secure Sockets Layer (SSL) or Transport Layer Security (TLS) to encrypt web traffic and prevent it being read or modified in transit.
- Use security certificates. With each page it sends, a web server verifies its identity by also sending the certificate. Browsers check the certificate's legitimacy to ensure the server is genuine.
- Use checksums. MD5 and FFP fingerprints are values (checksums) calculated using the data in a published document and embedded into the document. Publishers can give the checksum of a file they offer, and anyone can recalculate the fingerprint of the file they receive. If the calculated checksum does not match the published checksum, it proves that the document has been damaged or altered.
- Use email validation to verify that an email address exists and belongs to the correspondent.

Non-digital

- Compare original documents with any allegedly accurate copies. Contact the original authors of documents to verify their authenticity, if that is possible.

9780170440875

Relevance

Relevance measures how closely a resource, such as a book, database or web page, corresponds to people's desire for information.

TABLE 3.4 Circumstances when your data is in danger of becoming less relevant

Reason	Example
It is on a different topic. Using internet research to drift from link to link may lead to finding information that is irrelevant to the original topic.	A discussion of Scottish history mentions the sport of caber-tossing, which leads to a discussion of how the sport is scored, which somehow drifts into detailed coverage of kilts.
It is from a place where conditions are not comparable.	Comparing gun ownerships in suburban Australia with that of wild Canadian bear country.
It is from a different time.	'Boys in the 1950s were happy to leave school at 15 and get married at 18. What's wrong with kids nowadays?'
There are significant differences in history, conditions or circumstances that prevent two data sets being compared.	Cultural differences, war status, climate, characteristics of interview subjects and unusual recent events.

Accuracy

There are two main characteristics of data accuracy: content (functionality) and form (appearance). Content is divided into two parts: correctness and completeness.

Correctness

Correctness means that the values stored for a given object must be correct.

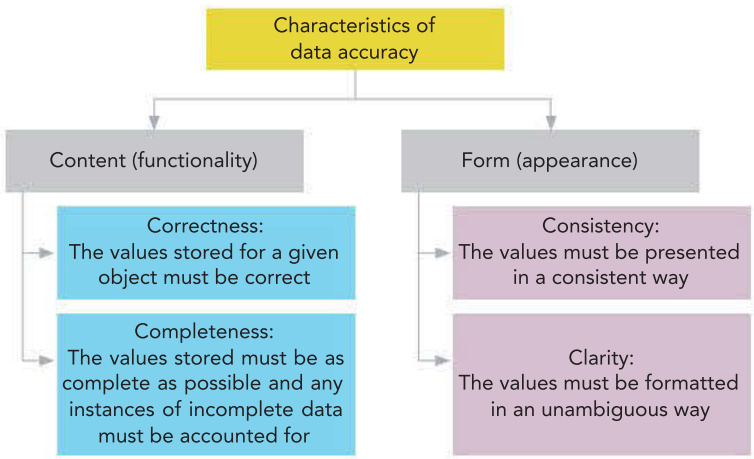


FIGURE 3.25 Characteristics of data accuracy

9780170440875

Faulty data acquisition equipment (such as inaccurate surveying software) and poor acquisition procedures (such as ambiguous questions) will pose a problem and introduce an accuracy problem from the moment you begin collecting data.

While you acquire your data, you may find that respondents misinterpret questions and answer inappropriately; for example, you ask a question about eating meat, and they assume the question excludes fish, and answer it incorrectly. They may also innocently misremember facts.

However, at times, inaccuracy is caused by genuine deceit – in the form of fraud, vandalism or criminal deception. Sometimes, it is less clear cut than that. Bias creeps into data if it is from a pressure group, which is an organisation driven by morals, special interests, profit, or even manipulative governments with a particular motive. If data providers have a **vested interest**, this should be acknowledged. For example, sometimes journalists need to declare that their travel expenses have been paid by the company whose products they are discussing.

Most research is based on samples from a much larger population. If the sample size is too small, this can mean the data is unrepresentative of the entire population. Sometimes the data collected is inaccurate because it is not detailed enough to provide a true answer; a pie graph alone does not offer enough detail for a serious statistical study.

Accuracy may be lost through data entry errors, such as simple typing mistakes. Data may also be distorted through improper statistical methods or poor language translation. Data may be stored, processed or communicated poorly.

Accuracy may also be lost as time passes. Changes in the real world since the data was initially collected may have a serious impact on the accuracy of the data. An overnight price rise can instantly make yesterday's advertisements or cost estimates inaccurate. Some data is dynamic (constantly changing), so if it is not maintained or synchronised, it becomes outdated. Most data needs to be updated, synchronised and have the necessary corrections applied over time to keep it current.

You can perform data quality assurance to cleanse or scrub data. This will identify and remove or repair data that is incomplete, inaccurate, irrelevant or inconsistent. It may also standardise data – for example, by changing all instances of 'Street' to 'St' or adding data to existing records, such as Medicare or tax file numbers.

Recalculate digital signatures or fingerprints to check if digital files have been modified or suffered from disk rot. You should also frequently synchronise copies with source data and conduct a quality check on secondary data obtained from a third party.

Completeness

Completeness means that your data set is just that: complete. All the values stored must be as complete as possible. Any instances of incomplete data must be accounted for.

Sometimes the data needed is no longer available because the records may be missing or the files have been lost as a result of fire, flood or war. Data from certain time periods may be unavailable. Digital data can be lost by accident or equipment failure. Disk rot or link rot may occur, or the format used to store the data may become obsolete. Obsolete data itself may be deleted rather than archived. Valid data may be removed accidentally during **data cleansing**.

In some circumstances, the data exists, but is inaccessible. For example, the data may have been collected but never published, or it has been censored by a court, privacy legislation, confidentiality agreement or oath. Sometimes data is deliberately ignored or buried because it does not support a particular answer to a question. For instance, counter-evidence may be ignored because researchers wish to show data from only one side of a debate.

9780170440875



Mark Fergus

FIGURE 3.26 Data may be incomplete because of water damage to paper records that have never been transcribed to computer.

THINK ABOUT DATA ANALYTICS 3.6

1 Identify the data formats shown in Figure 3.27.
2 Explain how you would access data found on only one of these obsolete formats.



iStock.com/Lezh

FIGURE 3.27 Data may be incomplete because it is only available on obsolete data formats.

The *Official Secrets Act* in several overseas countries still suppresses information of some events from World War II.

SCHOOL-ASSESSED TASK TRACKER

☑ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

Cherry picking or selective editing of data occurs when some data is chosen while other data is omitted; for example, researchers have a question about courts being too soft on crime, and try to support a preferred answer using data showing crime rates rising from 10 000 in 2019 to 12 000 in 2020, but in doing this they choose to exclude data showing that the crime rate in 2018 was much higher – 36 000. This data was excluded because it did not suit the story of crime figures increasing.

Simple error also introduces a great deal of incompleteness. Data entry errors occur. Researchers may fail to acquire the data for a relevant field; for example, forgetting to ask about alcohol intake when interviewing people about weight loss. They may ask irrelevant questions, such as asking single people how long they have been married. Sensors, both human and digital, can fail or measure inaccurately. If a database has been set up with overly stringent validation rules, it can prevent the input of unusual but valid data; for example, a 12-year-old genius cannot enrol at a university because its database refuses to accept the 'unreasonable' age of 12.

In some cases, respondents fail to answer certain questions. Sometimes they may refuse to answer sensitive questions. Participants may also drop out of studies before they are complete, leaving gaps in the research. Regardless of the cause, incomplete data can be just as misleading as incorrect data. Try to avoid drawing conclusions from incomplete data, particularly if the missing data could be vital to supporting or refuting your hypothesis.

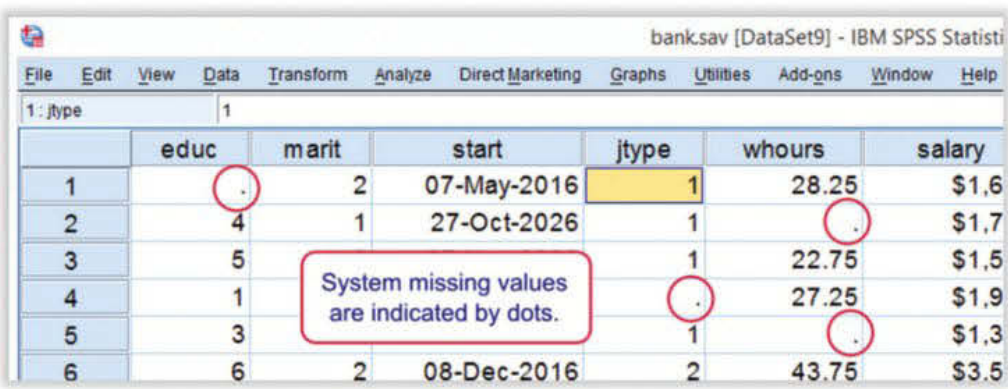
For data that you suspect has gaps because it has been collected but not published in full, you could contact the original data collectors to ask if they have unpublished data that they could provide you.

If the topic is controversial, look at data from proponents of both sides of the argument to see what inconvenient facts may have been ignored. Look for the hidden counter-evidence and find what has been cherry-picked. Researchers are often keen to point out the faults in evidence used by their competitors.

To minimise the potential for data entry errors, use existence validation to ensure that essential fields cannot be left empty when a database record is being entered. You can still use an incomplete record if it has valid data for a particular question, but ignore it for questions where it lacks relevant data. A single person who left the 'How long have you been married?' box empty could be used for calculating the number of times people get married, but ignored when calculating the average length of a marriage.

Validation techniques are also covered in *Applied Computing VCE Units 1 & 2.*

You can also interpolate the missing data. Interpolation is using a trend to estimate a missing value within a data set. For example, 2016 had 1000 sales, 2017 had 1100 sales, 2018's data is missing and 2019 had 1300 sales. If you interpolate 2018's missing data, it would have had approximately 1200 sales.



SPSS Tutorials, https://www.spss-tutorials.com/spss-missing-values/

FIGURE 3.28 Missing data usually means trouble for researchers.

When attempting to deal with completeness of data, consider the following.

- Avoid dubious methods of dealing with incomplete data.
- **Do not** use mean substitution to fill in missing data with the average value calculated from other records. Instead, use imputation to fill in a record's missing values using data from records that are otherwise similar to it.
- **Do not** simply ignore incomplete records during processing. This is just as troublesome as ignoring missing fields. The data may be incomplete for highly significant reasons, so removing the records could introduce bias into the information generated from the records and skew the results and conclusion.
- Be transparent. If your data is incomplete, do not try to hide it or ignore it. Sometimes incomplete data is unavoidable, so acknowledge it, explain what was missing and how you dealt with the issue.
- Make sure you investigate reasons for data being missing not at random (see Table 3.5). Data missing not at random may be significant to your research.

TABLE 3.5 Discussing incompleteness – some researchers discuss the data missing from their analyses, and apply categories to it. If you have data missing in your Outcome, you may wish to apply similar categories.

Category	Explanation
Legitimately missing	It is okay for it to be missing; for example, home phone number for a person who does not have a landline phone, years married for an unmarried person or hair colour for a bald person.
Illegitimately missing	This data should be complete, but is not.
Missing at random	There is no pattern to the missing data.
Missing not at random	There is a pattern to the missing data; for example, most men answered this question, but many women did not. This type of missing data should not be ignored.

THINK ABOUT DATA ANALYTICS 3.7

A real-life study ignored all records with missing data, and cherry-picked the data to use only records that were complete.

It later reported results saying that those who drink more alcohol have fewer problems with depression and anxiety, which contradicted common sense and decades of other research.

After appropriate handling of missing data, new results arose that were more consistent with the literature.

Suggest another case where the removal of incomplete records could lead to wrong conclusions.

Clarity

The form of data is important as well as the content because it will remove ambiguity about the content. Accordingly, form is divided into two parts: clarity and consistency.

Clarity is about formatting data in an unambiguous manner to prevent misinterpretation. For example, you have entered the dates of birth of all your interviewees into your database and they are the correct values. They are still not completely accurate, because it appears that some of the birth dates have been entered using the US date format (MM/DD/YYYY), and others using the Australian date format (DD/MM/YYYY). Worse still, some of the dates have both days *and* months under 12, making the actual interviewee subjects' dates of birth ambiguous. You can use an input mask to force a particular data entry format, such as XXXX-XXXX-XXXX-XXXX for credit card numbers.

YYYY-MM-DD is the international standard (ISO) format for dates. 2020-04-01 translates to 1 April 2020.

9780170440875

Is this date ...	
04/11/1996	November 4th, 1996 or 11th April, 1996
12/02/1999	February 12th or 2nd of December
06/09/2001	September 6th or 9th of June

FIGURE 3.29 These birth dates are inaccurate because you cannot tell what the intended or correct values are. You need to be explicit when entering data. In doing this, you can apply consistency and embed trust in the data.

Data entry errors and using multiple data entry operators can introduce ambiguity into the formatting of the data you have collected. Using multiple tables with different rules for formatting dates or numbers, such as dates of birth or calculated ages, can also pose a problem. Enforce data formatting and validation rules in your spreadsheets and database that prevent misinterpretation of the data. Try to avoid rekeying data where possible – copy and paste or import instead. Try out your questions with test subjects, such as your classmates or family, before conducting your research, so you can check that any interview questions or questionnaires make sense and are not ambiguous. Put dummy data in your database or spreadsheet to check for errors and potential problems with ambiguity before you begin inputting your real data so that you can troubleshoot. Do this for more than just one row or record – input a few so that you can test properly.

Consistency

Correct, unambiguous data can still cause a problem in a database if it is not consistent. Inconsistent data is unwelcome because it means the data is unreliable. This is why consistency is part of data accuracy in this Outcome.

Consistency problems can occur on several levels. For example:

- the date of birth listed in one table does not match age calculated in another table
- a respondent's answers to one survey question conflict with another
- data stored in one location, such as a local database, does not match corresponding data in another location, such as a linked website database.

You may cause consistency problems with your data if your formulas or queries have errors that calculate ages incorrectly or there are mistakes in data entry. You may also cause consistency problems with your data if you save multiple versions of your files without carefully managing them to avoid version issues. Your respondents may cause consistency

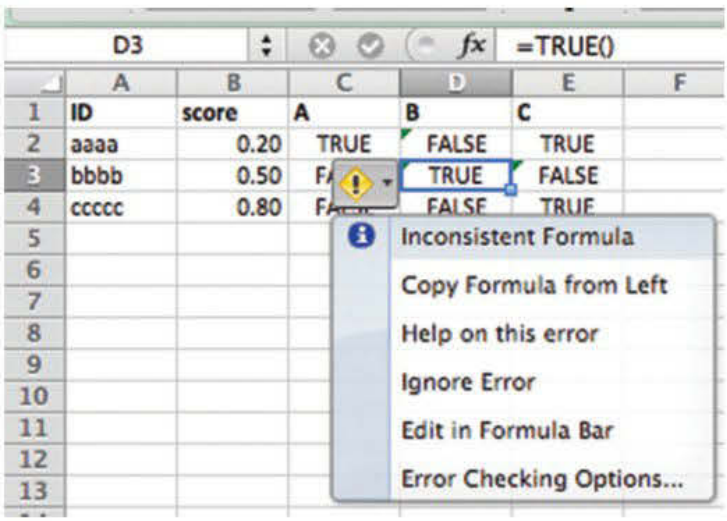


FIGURE 3.30 Microsoft Excel can detect inconsistent formulas and warn the user.

9780170440875

problems if they are not completely truthful or misinterpret questions. This means that consistency does have a degree of overlap with correctness, and therefore, functionality. When the data is inconsistent, it becomes hard to tell which is correct, and therefore, accurate information cannot be generated from the data.

When formulating questions that deal directly with people, ask the same question again in different ways to check the answer for consistency. This is important for questions about sensitive issues, because respondents may lie, avoid or bend the truth. People may manipulate their answers for any number of reasons, from embarrassment to exaggeration and simple memory lapse. If you have a research question that requires you to collect data about gambling habits, consider inserting 'Do you gamble?' at question 7 and then perhaps 'Last year, how often did you bet on a horse, buy a lottery ticket or use slot machines?' at question 38. People may forget they fibbed at question 7 by saying no and tell the truth on question 38.

You can later use electronic checking to detect inconsistent answers. For example, if your questionnaire asked how many children someone had in one question, and the age of their oldest child in another, the following would detect inconsistent answers:

```
IF (NumberOfChildren = 0 and AgeOfOldestChild > 0) then Consistency
Error
```

To prevent inconsistency caused by data entry, enforce consistent data formatting and validation rules. Try to avoid rekeying data where possible. Instead, copy and paste or import it after it has been initially keyed in. This advice is specifically repeated because it is important for data accuracy – you should try to only key data in once.

However, if some data *must* differ, use **fuzzy logic** to handle trivial differences in data. For example, accept 'Robert' if the user is also known as 'Bob'. Treat 'St Kilda', 'St. Kilda' and 'Saint Kilda' the same way.

Enforce referential integrity in databases to ensure key fields cannot be deleted without also deleting or modifying related fields. For example, in a relational database, it is important to ensure that if sections are deleted that any data linked to those is also deleted or moved to ensure they do not become meaningless. This was covered in Chapter 2. Any data stored in more than one place immediately invites the risk of inconsistency.

Timeliness

Data must be timely for it to produce usable information. That means it needs to be processed while it is current, and there should be no significant delays in retrieving it. Methods used to process data should be efficient enough to be complete by the time the data is actually needed. Make sure that the digital systems (hardware and software) used are powerful enough and appropriate for the task at hand to avoid causing delays. It is crucial that decision-making is never based on outdated data.

Protect users of your data as much as possible from potential delays caused by power outages, hardware failures, downtimes due to system upgrades, and deliberate threats from outsiders, such as denial-of-service attacks or malware infection. Make sure the age of data is known before using it: it may be meaningless to draw conclusions from very old data; for example, using data about food consumption that is 50 years old to draw conclusions about dietary preferences today.

THINK ABOUT DATA ANALYTICS 3.8

Why do you think there is an ISO standard for date formats?

Standardise data so it is always in a consistent, predictable and comparable format:

- Store family names in uppercase so that, when running comparisons or queries, 'Smith' is not overlooked as a match for 'smith' or 'SMITH'.
- Strip spaces from phone numbers at the start and end of all text data.
- Always use the same format when dates are stored as text.
- Use a standardised time setting, such as Greenwich Mean Time (GMT) or Australian Eastern Standard Time (AEST).
- Store all currency values in the same units: AUD, USD or GBP and so on.

The domain name system (DNS) is the biggest distributed database on the planet and lists every internet domain (for example, fred.com) and its IP address (for example, 192.34.78.101) by pointing to the location of the web server hosting the site's files. Changes to a site's IP address can take 24 hours to propagate (spread) throughout the DNS, so some people may quickly find the new home for a site while others receive 'file not found' (404) error messages.

9780170440875

Digital systems

While developing your SAT and throughout the course of your Data Analytics studies, you must use components of **digital systems**. The following sections discuss the composition of digital systems in terms of roles, functions and characteristics.

First, we begin with the main components of any digital systems, which are hardware, software and communications. Each component has a specific role to contribute to achieving the purpose of the digital system. Next are the functions of the components to support input, output and storage of data and information. Finally, the characteristics of each component will be considered.

Hardware

Hardware is made up of **input hardware**, which enables users to input (enter) data or provide commands to software, **output hardware**, which gives information to the user, **storage hardware**, which stores data and software, and **communication hardware**, which is any device that transmits a signal across a communications system.

Input and output hardware

The commonly used types of input hardware are described in Table 3.6.

TABLE 3.6 Types of input hardware

Hardware	Description
Keyboard	Used to type commands; Australian keyboards typically use the familiar QWERTY layout
Mouse	Used for controlling a graphical user interface (GUI) that has windows, icons, menus and a mouse pointer
Trackpad	Credit-card sized replacements for a mouse; sensitive to touch and taps and found beside or below the keyboard of laptop computers
Touchscreen	Common on phones, tablets, laptops, supermarket self-checkouts, ATMs and information kiosks; functions as both monitor and mouse replacement
Flatbed scanner	Digitise printed (analogue) text and images
Barcode reader	Used in shops and industry to increase the ease, speed and accuracy of entering product identification codes
Graphic tablet and stylus	Replaces a mouse for inputting intricate hand gestures for graphics, symbols, handwriting and signatures
Custom input devices	Includes buttons, knobs, switches and dials on game controllers; ticket machines; vehicle dashboards; music players; calculators; phones; remote controls and electric kitchen devices

The most common types of output hardware are monitors and printers. Monitors and printers have both undergone significant technological advancements across the years. Some of the most popular types of monitors are described in Table 3.7. The most commonly used printer types – laser, inkjet and thermal – are described in Table 3.8.

9780170440875

TABLE 3.7 Popular types of monitors

Monitor	Description
LCD Liquid crystal display	• Thin • Colourful • Uses comparatively little electricity • To be visible, LCD pixels (picture elements) need a backlight
OLED Organic light-emitting diode	• Pixels glow to produce the display • When turned off, a pixel is black, which produces greater contrast and richer colours
Plasma	• Pixels glow and are truly black when turned off • Uses more electricity than LCD • Cheaper than LCD for very large screens and offers richer colours and better contrast
Electronic paper	• Electrophoretic ink (for example, e-ink) • Requires no electricity to hold a text or image display once it has been drawn • Extremely low power consumption • Ideal for battery-powered devices with large displays that are turned on for long periods (such as ebook readers)

All printers tend to include lights, dials and indicators to signal activity or status. They also tend to use speakers for alarms and sound effects when there is a problem, or to indicate completion of a task. You have likely often heard the beep your printer makes when it will not print because of a paper jam or because it is out of ink or toner.

TABLE 3.8 Laser, inkjet and thermal printers

Printer	Description
Laser	• Fast, high-resolution, waterproof pages • Monochrome or colour text and graphics • Uses a laser beam to draw an entire printed page onto a rotating drum; the drum passes over ultra-fine plastic powder (toner) and is drawn to the static electricity left where the laser beam hit the drum; the drum transfers the loose toner powder onto a sheet of paper and a heater melts the toner onto the paper
Inkjet	• Sprays ink onto a page one line at a time • Cheap to buy • Ink is very expensive
Thermal	• Used in industry for labels and barcodes, and in retail for receipts • Operates in near silence • Uses very little electricity, which makes them the best choice for battery-powered portable printing, such as for printing parking fines • Printing heads burn heat-sensitive paper to leave text or images • Text and images will fade after a relatively short time, which makes them not a good choice for longevity

THINK ABOUT DATA ANALYTICS 3.9

Explain conditions where laser and inkjet printers are either ideal or unsuitable.

Storage hardware

Storage hardware stores data for immediate or later use. Storage hardware, therefore, is made up of primary (short-term) storage, which you know as random access memory (RAM), and secondary (long-term) storage. Secondary storage includes many kinds of disk drives.

9780170440875

RAM is composed of billions of storage locations in silicon chips. RAM stores program instructions and data when programs are running. RAM chips are volatile and lose their data when power is turned off. Typical computers now have 16 or 32 GB of RAM at a speed of 2400 MHz DDR4.

Secondary storage options permanently store data, information and applications when they are not being used. This includes when power is turned off. Table 3.9 discusses three of the most popular secondary storage options: hard disk drives, solid state drives and network attached storage.

THINK ABOUT DATA ANALYTICS 3.10

What was a typical RAM size and speed for early computers? What is the impact of increasing the amount of RAM available, and the speed of the RAM?

TABLE 3.9 Popular secondary storage devices

Storage	Description
Hard disk drive (HDD)	• Aluminium disks densely crammed with magnetically recorded bits of 0s and 1s (binary code) • Spinning at up to 10 000 RPM, they store and retrieve data at incredible speed, with breath-taking accuracy and reliability • Very cheap per megabyte of capacity; still the largest, most reliable long-term storage media • In 2018, a 6 TB (6 terabytes = 6000 gigabytes) HDD cost approximately $AUD200 – around 30 GB of storage per dollar
Solid state drive (SSD)	• Non-volatile memory similar to USB flash drives and SD cards • Runs silently, starts instantly, generates less heat and uses less electricity • Has no motors that will age and eventually fail • Tends to access data faster than a HDD • Stores less data per square centimetre of storage space • Has a limited number of times it can rewrite a memory cell; after approximately 1 million writes, a memory cell will become unpredictable, or fail • Quite expensive: in 2018, a 500 GB SSD cost approximately $AUD200 – around 2.5 GB of storage per dollar
Network attached storage (NAS)	• A networked team of HDDs in a box • Makes file sharing easier • Increases capacity considerably (for example, 12 TB) • Offers data protection, such as hot-swap disks • Convenient and reliable

TABLE 3.10 RAM and data storage units

Unit	Symbol	Equivalent to:	
		RAM	Data storage
Byte	B	8 bits (0 or 1), the basic unit of storage	8 bits
Kilobyte	KB	1024 bytes	1000 bytes
Megabyte	MB	1024 KB	1000 KB (roughly 1 million bytes – the size of two average novels)
Gigabyte	GB	1024 MB (PCs have gigabytes of RAM)	1000 MB
Terabyte	TB	1024 GB	1000 GB (hard disks have terabytes of storage)
Petabyte	PB	1024 TB	1 000 000 GB
Exabyte	EB	1024 PB	1 000 000 000 GB
Zettabyte	ZB	1024 EB	1 000 000 000 000 GB
Yottabyte	YB	1024 ZB	1 000 000 000 000 000 GB

For more information on input, output, storage and network hardware, see Chapters 3 and 7 of *Applied Computing VCE Units 1 & 2*.

9780170440875

Network and communication hardware

Networks and communication hardware come in a variety of forms. Ports are the physical sockets that carry data between a computer and peripheral (external) devices. For example, USB (universal serial bus) is an industry standard high-speed protocol that connects many devices, such as flash drives, printers, modems, keyboards, mice, speakers or phones.

Modems convert digital data into analogue data for transmission over non-digital media, such as telephone lines, and convert incoming analogue data into digital data for the computer to use. The two most common types of modem are ADSL and cable. ADSL is short for asymmetric digital subscriber line; asymmetric is a reference to the fact that the modem downloads (receives) data faster than it uploads (sends). Cable modems use pay TV cables, such as Foxtel and Optus.

Switches are devices that allow multiple network cables to interconnect and exchange data between networked devices.

Routers are used as gateways to interconnect LANs and they guide data packets across networks and the internet. At home, your router is normally built into your modem and it protects your LAN from the outside world.

Cables are used to connect two devices, enabling the transfer of signals from one device to another. The most common networking cable is CAT, a metal-core twisted-pair cable, usually labelled with a number (for example, CAT 6). Its maximum length is 100 metres, and its maximum bandwidth is 1000 megabits per second (Mbps). For longer distances and higher speeds, glass fibre-optic cable (FOC) is used. FOC can run for kilometres, is immune to electrical interference, runs at near light-speed, and has massive bandwidth because, unlike CAT cable, it can carry multiple signals along a single thread. FOC is used to span oceans and continents.

Wireless access points are devices used on wireless LANs. They act as central transmitters and receivers of wireless radio signals and allow wireless devices such as phones or tablets to connect to the wired network. Although wireless is cheap to install and flexible to configure, it is slower and less secure than a wired connection. If you use wireless data, you should encrypt your data to safeguard your personal and sensitive information.

Software

Software is the programming code that controls hardware. Software comprises:

- application software – for example, a word processor – to do work for the user and create information
- system software – for example, an operating system or device drivers – to control hardware and allocate computer resources so application software can run
- utility software – for example, a text editor – to provide a single, specific service to extend the functionality of a digital system.

Software is written in source code that uses the strict syntax of a particular programming language (for example, C or Basic C#, Visual Basic, JAVA, JavaScript, PHP, Python, Wolfram Language, and so on) that is converted by compiler software into executable code (for example, notepad.exe) that a computer's central processing unit (CPU) can understand and carry out. We have already discussed spreadsheet software, such as Excel, which is relevant to your Outcome. Databases such as Microsoft Access, LibreOffice BASE and FileMaker, and word processors such as Microsoft Word, have also been mentioned.

9780170440875

GanttProject
GanttPRO
ProjectLibre

Gantt charts

GanttProject is downloadable, free, open-source software. All of the Gantt charts used in Chapter 3 were made using GanttProject. Another open-source option is ProjectLibre. There is also Microsoft Project, which comes with the professional version of the Microsoft Office suite. If you wish to use web-based software, GanttPRO is free for basic or personal users.

GIMP
Balsamiq

Editing images and creating artwork

To edit images or create new artwork, you could use Adobe Photoshop and Illustrator. For an open source, free option, you could download GIMP. Microsoft Word has a 'Smart Chart' feature that can create hierarchies and flowcharts and other types of charts. It is customisable and simple to use. To create mock-ups and annotated diagrams, Balsamiq is free for a 30-day trial usage or simply use PowerPoint or Keynote to generate a .jpg file.

SurveyMonkey
Typeform

Survey tools

Google Forms, SurveyMonkey and Typeform provide browser-based, free, easy-to-use and easy-to-distribute templates for making questionnaires. Responses to questionnaires are simple to access and you can set up alerts to be notified when new responses are received.

Tableau Public
Inspiration

Data visualisations

Most browser-based data visualisation tools are offered on a trial basis: this includes online tools such as Lucidchart, while Piktochart is free for education. There are exceptions to this, of course, but you may need to look around for the latest free tools that best suit your needs. Google Charts and Google Studio are also offered within the Google suite of free tools. To create mind maps of your data, you could download a trial of Inspiration. Another possibility is a trial version of Tableau Public to visualise your data for free – but do be aware this may also *share* your data to a public gallery as a condition of use. Alternatively, your teacher could apply for an education licence, which is valid for up to a year, and this would provide a registration code to convert the trial to a full version. Mathematica is available to all students and teachers from all sectors at no cost.

Free Mathematica licence

Data security

Key legal requirements for the storage and communication of data and information is covered in full in Chapter 7 on pages 304–14.

As part of Unit 3, Outcome 2, you must select and apply methods to secure and store data and information. Losing the data that you have painstakingly gathered for your Outcome would be catastrophic. To a business, the loss of data, depending on magnitude, can be anywhere from minor to fatal. For example, if a company lost records of accounts payable, it would not be able to keep track of what had been paid and what still needed to be paid. If a retail store lost track of stock lists, it would have no way of knowing if stock levels indicated a great deal of shoplifting had been taking place, or if staff had been making mistakes when putting transactions through.

Businesses may lose trade secrets to competitors and lose their reputations as trustworthy organisations if they fail to protect their data and information. They also face prosecution by the Australian Tax Office if their tax records are lost, as well as prosecution under the *Privacy Act 1988* (see page 307) if they violate the Act and personal information is lost, damaged or exposed as a result.

9780170440875

In addition, they may lose income if business cannot carry on because of the loss, and may find they are unable to pay wages.

Clearly, it is very important that businesses protect personal and corporate data from accidental or deliberate loss, damage or theft. This is an ongoing responsibility for them. However, it is vitally important for you as well. Although it may only seem relevant for your Outcome, developing good habits for data security right now is something that is worthwhile for the long-term.

Just as businesses can lose their data and privacy, you can lose yours. Unfortunately, malicious behaviours such as identity theft and **doxing** do happen. There are also a multitude of other data security threats that are not malicious that are nevertheless real and dangerous. Regardless of how unlikely it may seem, it is best to employ data security measures. Data security takes two main forms: physical and software. The next section discusses the physical and software security measures relevant to you and your data.

Doxing involves researching a person, sometimes known only by a handle (nickname or screen name) and then publishing their personal information, such as their full name, address, phone number, workplace and date of birth, online to identify them to as large an audience as possible.

Physical security

Physically keeping threats away from your data is a logical and effective first step.

- If you use a laptop or tablet and store your data on it, make sure you keep it in a secure place when you are not using it, such as in a cabinet. Otherwise, store it out of sight.
- Do not let people you do not know very well use your devices or guest accounts, and avoid providing user or admin access to your device.
- If a friend or family member needs to use one of your devices, make sure they cannot access important data. Log out of social media and email to avoid posts being made without your knowledge.
- Keep your doors and windows locked to prevent theft of your hardware.
- If you use a desktop computer, keep it switched off when you are not using it.
- Consider using surge-protector power outlets for all of your devices to protect the data stored on them.

Software security

Software security is extremely important to you and your data. Threats can come from anywhere at any time, and they can be both deliberate and accidental. Threats can come across a network and do not rely on physical access to equipment.

Use strong passwords

Ideally, passwords should be at least eight characters in length and include a combination of both uppercase and lowercase letters and numbers. Depending on the software, you may also be able to include special characters such as punctuation. Do not use the same password on every login. If you do, and a hacker gets into one of your accounts, they will then have access to *all* of them.

Passwords that contain common words are easily guessed, so are not recommended. Examples of password strength are shown in Table 3.11.

9780170440875

How secure is my password? These websites rate the entered string – never enter a 'real' password.

How secure is my password?

Kaspersky Secure Password Check

The Password Meter

TABLE 3.11 Strong versus weak passwords

Weak passwords	Strong passwords
Sunshine1	tYjL3!1pC
nothing00	N0tS4y1nG!
password	P4s3w0Rd?!
qwerty123	PinkHorseJumpFast (strongest password)

It is also a good idea not to use the same secret question information on every account. While it may be convenient to always use your mother's family name or the name of your first dog as your 'Forgot password?' reminder, if a hacker finds this out, it will make it very easy for them to take over every single one of your accounts using the 'Forgot password?' feature.

For some logins, you can also use two-factor authentication. For instance, a login to myGov.au requires a password, followed by the entry of a secret code that has been sent to the user's mobile phone. Two-factor identification relies on identifying people because they both know something (a password) and possess something (such as a mobile phone).

Use login passwords

You should use login passwords on your laptop, tablet, desktop computer and any other electronic device that has personal or sensitive information on it. You do not want to risk losing it or having it stolen, and having someone else switch it on and have immediate access to everything you have stored on it.

Use biometric identification

While passwords are the only way to control access to remote computers and resources (currently), biometric identification can be used to control computers and resources when the user is physically present. Biometric data cannot be lost, stolen, guessed or discovered easily. Biometric signatures are unique, and include fingerprints, iris patterns (the coloured part of the eye) and retinal patterns (the blood vessels at the back of the eye). Biometric identification has long been used at Los Angeles airport. For years, non-US citizens underwent fingerprint scans to enter the country. Now, the airport is considering an upgrade to iris scan technology. Facial recognition is also emerging as a unique identifier – this technology has been trialled for reliability at several Australian airports and is expected to eventually replace boarding passes and passports.

Australian airports prepare for facial recognition (Qantas)



FIGURE 3.31 While most types of biometric data are not yet relevant to you for personal use, the following may be more relevant: **a** iris pattern, **b** facial recognition, and **c** fingerprints.

9780170440875

Some laptop manufacturers, such as Acer and Toshiba, sell laptops with built-in biometric identification systems in the form of fingerprint scanners. If you have a laptop with a fingerprint scanner (Figure 3.32), scan your own fingerprints and use them as your default login for added security.



Getty Images / iStock / rusm

FIGURE 3.32 The fingerprint scanner positioned directly below the trackpad (in-built mouse replacement) on a laptop computer. The user only need run their fingerprint over the brown stripe in the gold section to capture their fingerprint scan.

Always log out

When you are not actively using a computer, do not leave it logged in. Log out and turn off the monitor. Leaving the computer logged in to your user account leaves your data vulnerable to anyone who walks past and sees that the computer is still logged on.

Log out of websites when you are not using a personal device. In fact, it is better to log out of websites even when you are using your own personal devices as well, but people rarely do this, which makes it easier to steal data and identities from people whose devices have been misappropriated.

Imagine if your phone was stolen. You left yourself logged in on Dropbox on your phone, which was storing files from your Outcome. The thief decided to clean out the phone and wanted to see if there was anything of value on Dropbox. Having no interest in Data Analytics, all the files were deleted. Next, they wanted to see what you had in your email, so they went into your email account, which was logged in, and changed the password. It was easy to do this, because your mobile phone browser gave you the option to view your saved passwords. Knowing it would be easy for you to get control of your email account back, they changed your 'Forgot password?' questions.

Perusing your email inbox, they established that you had not deleted your monthly bank statement email. It included your bank account number. Suspecting you were a creature of habit, the thief went to the bank's website and logged in using the same password as your email account. Using a few quick steps, a thief can access your bank account.

You can prevent all of this by not storing passwords in your browser, always logging out, and not using the same password repeatedly.

Encrypt your data

Always encrypt your data. This is especially true if you use wireless connectivity at home. Wireless data is more vulnerable to threats than wired data. Encryption makes data unreadable to unauthorised people, even if they manage to overcome other security measures and steal it.

THINK ABOUT DATA ANALYTICS 3.11

Breaking a 256-bit key by brute force (simply trying every possible key in turn) is not simply twice as hard as cracking a 128-bit key; it is 2^{128} harder.

To help you visualise that value:

- 2^2 is 4 times harder.
- 2^{10} is 1024 times harder.
- 2^{20} is more than a million times harder.
- 2^{21} is more than 2 million times harder, and so on.

1 Exactly how much harder is 2^{128}?

2 Calculate how many billions of years it would take to crack a 256-bit encryption key using 50 supercomputers. In November 2018, the fastest supercomputer, Summit, achieved a benchmark speed of 144 petaflops. A petaflop is 1000 trillion floating point operations per second (10^{15}) Confirm the latest supercomputer processing speed.

9780170440875

Modern encryption uses **public key encryption**, which does not need a key to be sent to unlock it. Previous encryption methods required a secret unlocking key to be sent with the encrypted data. If the data was intercepted, the key could also be captured and the data unlocked.

Public key encryption is the basis of SSL and TLS, which are used to encrypt the web traffic between servers and browsers. If intercepted in transit, the traffic is unreadable. **Pretty Good Privacy (PGP)** is software that also uses public key encryption to protect documents. Wireless signals also use public key encryption to prevent snooping and unauthorised use of wireless networks.

The bigger the numbers used for public key encryption, the harder the encrypted data is to decode. For greatest safety, choose the largest encryption key you can (128-bit or 256-bit is the current recommendation).

Use a firewall

A firewall will prevent unauthorised access to your data and information, and deny network access to outsiders. Essentially, it will separate the internet and other networks from the computer or LAN on which it is installed. A firewall examines the content of incoming data packets and determines whether they should be allowed to pass through.

ZoneAlarm

High-quality free firewall software includes ZoneAlarm, Comodo Firewall, PeerBlock, Privatefirewall and Anti NetCut 3.

Use antivirus software

Most antivirus programs handle a whole lot more than just viruses now. There are numerous kinds of malware around these days (Table 3.12), so most antivirus programs are equipped to handle at least a few of them.

TABLE 3.12 Common types of malware

Malware	Description
Viruses	• Damaging code that attaches to executable files and travels with them. • **Payload** is triggered by human actions, such as running programs. • True viruses are now rare, but the name persists as a generic term for malware.
Worms	• Copy themselves and travel with no human intervention or need to attach to other files. • Most travel via email or over local area networks.
Spyware	• Monitors user behaviour and reports browser activity to the spyware's operator. • Can hijack browsers to send users to unwanted sites or show targeted advertising based on a user's browsing history.
Trojans	• Any malware that enters a system by pretending to be desirable. • Often use 'social engineering' to trick people into downloading or installing them.

TABLE 3.13 Known malware payloads

Payload	Destructive potential
Keylogger	Can record users' keystrokes including passwords, credit card information and bank account logins, and send the data to the malware operator.
Distributed denial of service (DDoS) attack code	Makes a computer vulnerable to remote control, along with thousands of other infected machines, to participate in a DDoS attack on a remote victim. A computer that might seem sluggish to its user might actually be sending millions of information requests that can render the victim's computer unable to operate. DDoS attacks are often used to blackmail victims into paying protection money or to attack political or religious enemies.

9780170440875

Payload	Destructive potential
Adware	Inserts unwanted advertising into visited websites. Not deliberately destructive, but often poorly programmed and can dramatically slow computers down or cause crashes.
Spam server	Sends thousands of spam emails using the victim's computer. If discovered, innocent computer owners are identified while the spam operator remains undetected.
Ransomware	Encrypts documents on the victim's computer so they are inaccessible to the computer owner. The malware operator then demands payment from the victim to receive the key to unlock their documents.
Root kit	Particularly nasty malware that actively hides from the operating system and works invisibly in the background, often turning a computer into a remote-controlled 'zombie'.
Deleting files or damaging operating systems	Such petty vandalism, once common, is now rare. Cyber-attacks are now dominated by large, organised crime syndicates and governments.

While your home computer is unlikely to be deliberately targeted by international hackers who are determined to steal your valuable drafts of Data Analytics solutions, you do need to protect your computer system and online accounts. Having a sound understanding of electronic self-defence is necessary for both your digital system and your end-of-year Data Analytics examination.

Even the most careful computer user cannot guard against worms or 'drive-by downloads' that can cause infection simply by visiting an infected website. However, keep the following in mind.

1. You *must* use a reputable, reliable anti-malware scanner.
2. You *must* always have it running and scanning opened and downloaded files for known threats.
3. You *must* keep your virus definitions up-to-date.

Be aware of false positives. Sometimes a scanner can report a virus that does not exist. Free online scanners sometimes report false positives to scare users into buying their products.

Similarly, be aware of false negatives. Some scanners may be unable to detect existing viruses. This may occur if your virus definitions are out of date and with newly released 'zero day' threats.

If possible, have a discussion with your parents about paying for a reputable malware scanner to avoid false positives. (Some free online scanners, ironically, also have in-built adware.) If this is not an option, research carefully to choose an antivirus program that is right for you.

Backup your files

Despite all of your best efforts, disasters may still happen. Data backups are the final defence against total data loss.

Backup your files, especially any files you create for your Outcome (and your VCE studies in general), *at least once a day* depending on how often you are changing them.

9780170440875

THINK ABOUT DATA ANALYTICS 3.12

Conduct a web search for '3-2-1 backup strategy'.

1 What is the 3-2-1 backup strategy?
2 Will you adopt the 3-2-1 backup strategy?
3 Why/why not?

For some assignments every 10 minutes will prevent data loss should power fail during a long writing session. You cannot lodge a request for consideration if you lose your digital files. While a minimum once a day is recommended, it is better to err on the side of caution and backup every time you make a significant change.

Store backups away from your main device. While companies can store their backups off-site, the best you can do is probably store your backups on external drives as well as on your internal hard drives. Alternatively, consider backing up your files on the cloud using a service such as OneDrive (as part of an Office 365 subscription), iCloud (5 GB free), Dropbox (2 GB free) or Google Drive (15 GB free).

Test your backups to make sure they work. Also regularly create system restore points on your computer in case anything goes wrong and you need to restore to an earlier version.

Key legal requirements for storage and communication of data and information

State privacy legislation as well as the Australian Privacy Principles (APP) are fully detailed in Chapter 7 (page 309). These legal requirements include human rights, intellectual property and privacy. The Chapter 3 end-of-chapter questions 26–29 will refer to this material.

Next steps

In this chapter, we discussed the features of a reasonable research topic or question, and we also discussed data in many forms: primary and secondary, qualitative and quantitative, how to gather it directly, how to gather it using resources and how to reference it correctly in your report.

We then discussed the digital systems that you have been using throughout your studies: the hardware, software and networks enabling your data gathering and research.

Finally, we discussed data security, with regard to how it relates to you and your studies and the key legal requirements for storage and communication of data and information.

Your next step, upon completion of the chapter summary, is to work towards completion and submission of the solution for Unit 3, Outcome 2, according to your teacher's instructions.

As you collect data for your Outcome, take steps to protect respondents and subjects, maintain your data's integrity and apply appropriate data types and structures – while you do this, you should be thinking about relevant legal constraints. Chapter 4 begins with a discussion of the solution specification and design requirements for Unit 3, Outcome 2.

9780170440875

3 CHAPTER SUMMARY

Essential terms

analytic coding the means by which labels are applied to transcripts of text

cherry picking selecting data to serve preconceived ideas about the results; contradictory data is excluded from the sample

citation a reference to the information's source

closed question a question that offers a limited range of possible answers

communication hardware any device that transmits a signal across a communications system (for example, modem, router, wi-fi card, network interface card, and so on)

concept an idea

concurrent tasks in reference to project management, carrying out one task at the same time as another

controlled when variable is maintained unchanged, while other variables are changed

data raw, unprocessed facts and figures

data cleansing process where inaccuracies in data are detected and corrected either by changing, replacing or removing

descriptive coding summarising, in a word or noun, the basic meaning of a passage of qualitative data

digital object identifier (DOI) a unique alphanumeric string assigned by the international DOI Foundation; a persistent link to an internet location is assigned once an article is published

digital system hardware for input, output, storage and communication

doxing when a person who only provides an avatar and pseudonym is 'documented' by having name, address, IP address, phone number published online to expose their identity

event an activity of interest

fair test changing only one factor at a time in a test, and keeping all other conditions the same

frequency distribution table a table that summarises values and frequency in two columns

fuzzy logic lying between true and false

Gantt chart a chart that tracks the progress of a project by placing tasks on a timeline, often with comments or annotations

graph a visual representation of data showing the relationships between several elements

information useful knowledge created by manipulating data

input hardware equipment that allows data to be entered into a computer system, such as a keyboard, mouse or joystick

interviewer bias the interviewer's preconceived expectations of the results; the interview questions and answers confirm that bias

Likert scale a method of encoding qualitative data, usually on a numbered scale

mind mapping visually organising information into a diagram

output hardware devices that allow computer-generated information to be converted to human readable form, such as text, graphics, touch, audio and video

paraphrasing using another person's ideas but putting them into your own words, rather than directly quoting them

payload the cargo to be delivered

predecessor a task that must be completed before another task can begin

9780170440875

Pretty Good Privacy (PGP) an encryption program that provides privacy and authentication for data communications

primary data new facts collected personally by a researcher to answer a specific question

process a procedure or instruction that is followed

public key encryption data is transmitted confidentially using a combination of a publicly known key and a private key known only to your computer

qualitative data expressed in words; concerned with opinions, feelings and experiences

quantitative data data that is expressed as numbers, categories or labels and is easy to process

secondary data data collected by someone other than the researcher, which has often been processed

skewed result the frequency of responses by the population is not evenly distributed, rather it 'leans'

slack time the length of time that a task runs overtime before it affects other tasks

storage hardware the physical location for your files to be held for later access; this may now include cloud storage where the storage drives are remote from your computer device

successor task that must be completed after another task

survey distortion when survey results are influenced by external factors

theoretical coding *see* analytic coding

vested interest where an individual, group or organisation has an investment or interest in an issue and stands to gain or lose from it

work breakdown structure (WBS) breaking down your project into achievable tasks

9780170440875

Important facts

1. **Data** is made up of raw, unprocessed facts and figures.
2. **Information** is derived from processing data into a form humans can understand.
3. **Gantt charts** show the progress of a project by placing tasks on a timeline, often with comments or annotations.
4. There are many government and non-government **public sources of data sets** in Australia.
5. **Quantitative data** is expressed as numbers, categories or labels and is easy to process.
6. **Qualitative data** is rich, unstructured textual data containing opinions and must be encoded to be processed.
7. **Likert scales** are often used to encode qualitative data.
8. Qualitative data may use **rubrics** to guide how freeform ideas are converted to labels or numbers.
9. **Data types** include number, character (text, string) and Boolean (true/false). The data types of database fields must be chosen with care.
10. Use the American Psychological Association (APA) **referencing system** for your Outcome.
11. Proper referencing is needed to avoid **plagiarism**.
12. The **APA referencing style** inserts the author and year of a reference into the body text.
13. A **reference list** contains full details of every reference given in the body text.
14. **Personal information** can identify a person (for example, name, address and phone number).
15. **Sensitive information** includes data on medical history, politics, sex, religion, race and so on.
16. **Health information** includes medical history, diagnoses, drug prescriptions and genetic information.
17. Databases and spreadsheets must be **organised** to be effective.
18. Organise data with **sorting** and **filtering**.
19. Data visualisations make **data patterns and relationships** clearer than lists of numbers.
20. Data can be secured **physically** (for example, locked doors), with **software** (for example, passwords, encryption, access hierarchy) and with appropriate procedures.
21. **Backups** must be regular, stored off-site and tested. They are a key component of a 24-hour **data disaster recovery plan** that all organisations should devise and practise.
22. **Malware**, such as worms and viruses, can cause significant security problems and must be guarded against.

9780170440875

TEST YOUR KNOWLEDGE

Review quiz

What is data?

1 Identify the major difference between data and information.
2 Identify the main characteristics and strengths of primary data.
3 Identify the main characteristics and strengths of secondary data.
4 List five sources of primary data and five sources of secondary data.

Quantitative and qualitative data

5 How can qualitative data be prepared for statistical manipulation?
6 Give an example of a question using a Likert scale.
7 Explain the usefulness of a rubric.
8 Contrast qualitative and quantitative data.

Acquiring data

9 State reasons why you would avoid using loaded and biased questions.
10 Summarise the advantages of using interviews to collect data.
11 When is observation the most appropriate data acquisition method?
12 Explain one advantage that surveys and questionnaires have over observation and interviews.
13 Write a query that would find males between the ages of 16 and 18 inclusive, and females older than 25.

Referencing data sources

14 How does proper referencing avoid plagiarism?
15 Give examples of using parenthetical (author, year) and numbered reference styles in body text and reference lists.

Data types and data structures

16 Why is it important to choose proper data types?

Data integrity

17 Summarise the meaning of 'integrity of data'.
18 How does lack of timeliness degrade the value of data?
19 How does incomplete data lead to faulty information?
20 How does database normalisation contribute to data accuracy?

9780170440875

TEST YOUR KNOWLEDGE



21 Describe one way to discover if people are not telling the truth in surveys.

22 What does 'data authenticity' mean?

23 List three examples of how data may lose relevance.

24 What can be done to maintain the accuracy of data over time?

25 What is the main difference between open and closed questions, and when should each be used?

Legal requirements

26 Why is privacy legislation necessary? What are the consequences of not knowing or following privacy legislation and intellectual property laws?

27 Why should you cite your sources? What is the implication if you do not?

28 How would you reassure your interviewees that their responses would be kept private?

29 What steps can you take to safeguard the intellectual property you create for your Outcome?

File-naming strategies

30 a Write a brief list of file-naming rules suitable for a novice computer user.

b Add four file-naming conventions to this list.

Organising and storing data

31 List three ways to organise a spreadsheet effectively.

32 Explain two important factors when naming objects that will affect the usefulness of sorting.

33 How does synchronising improve the integrity of stored data?

Digital systems

34 Create a table listing digital system components for input/output, storage and communication. For each component, outline the main types that exist and their strengths and weaknesses.

Data security

35 Your family is concerned that the precious digital photos on your home computer may be eventually lost or damaged. List deliberate and accidental threats to these photos. Recommend physical and electronic strategies and techniques for protecting and recovering these irreplaceable files.

SCHOOL-ASSESSED TASK TRACKER

☑ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

APPLY YOUR KNOWLEDGE

1 Choose a short-term practice topic for research. Consider data sources that may provide evidence to refute or support your research question. Your classmates could be useful primary sources. Examples of potential topics that could be honed into a reasonable research question could be:
   - teenagers prefer one type of public transport
   - soccer is rising in popularity
   - your school compares well with neighbouring schools
   - internet shopping is better than personal shopping
   - any other operating system is better than Windows.

2 The nature of the research topic will determine the data types you need. You may consider creating a set of interview questions for a few (for example, three) of your peers to gather qualitative primary data about your topic. Other types of primary data may include photos and observation. For secondary data, you may need to access journals, newspaper articles, television reports, radio programs and statistical evidence.

3 Create a survey or questionnaire for a larger number of your peers on the same topic (for example, 10 of your peers).

4 Conduct the interviews and distribute the surveys or questionnaires. Record any problems you encountered acquiring this information. What questions may have needed changing?

5 Encode the interview data to let you summarise it into meaningful categories and trends.

6 Write a report on the research topic containing the following.
   a A statement of the research question
   b Summaries of interviews, including a copy of the questionnaire
   c Supporting quantitative secondary data from printed or online sources, fully referenced using APA style; primary data must also be referenced
   d A statement evaluating the quality and integrity of the data used to generate information about the research question
   e A statement explaining how the data will be stored securely
   f A statement about how the data was evaluated to generate information
   g A statement about the research question using information derived from your data

9780170440875

CHAPTER

4 Drawing conclusions

KEY KNOWLEDGE

After completing this chapter, you will be able to demonstrate knowledge of:

Approaches to problem solving

- functional and non-functional requirements, including data to support the research question, constraints and scope
- types and purposes of infographics and dynamic data visualisations
- design principles that influence the appearance of infographics and the functionality and appearance of dynamic data visualisations
- design tools for representing the appearance and functionality of infographics and dynamic data visualisations, including data manipulation and validation, where appropriate
- techniques for generating alternative design ideas
- criteria for evaluating alternative design ideas and the efficiency and effectiveness of infographics or dynamic data visualisations.

Reproduced from the VCE Applied Computing Study Design (2020–2023) © VCAA; used with permission.

FOR THE STUDENT

This chapter concludes the discussion of the theory and skills required for Unit 3, Outcome 2.

You will be introduced to solution requirements, techniques and tools for planning and project management, identifying patterns and relationships in data.

By the end of this chapter, you will be ready to frame a research question, gather data to shape your investigation, report your findings, and cite the data sources you used.

FOR THE TEACHER

This chapter concludes the theory and skills needed for Unit 3, Outcome 2. Having covered the theory of data in Chapter 3, students are now introduced to solution requirements, design principles, and techniques and tools for planning and managing the progress of a complex project.

By the end of this chapter, students should be equipped to devise their own research topic or question, search for relevant data from which they derive information to investigate their topic, and report their findings with evidence and a formal citation of sources. **Note:** You are not expected to provide students with a research topic or data.



Shutterstock.com/Nomadic Photo Studio

9780170440875

Continuing Unit 3, Outcome 2

In Chapter 3, you learnt how to research a reasonable, testable research topic or question. You also learnt about data: primary and secondary data, the difference between qualitative and quantitative data, how to protect data integrity and how to reference sources correctly.

These factors are all relevant in preparing for Unit 3, Outcome 2. For this Outcome, you will submit the information you have produced from the data you have gathered, which attempts to answer your research question.

We begin by discussing how to narrow the scope of a solution to identify the relevant constraints. Next, we will cover setting the specifications for a solutions both in terms of a general solution and your specific solution.

We will also talk about identifying patterns and relationships in data, because this will help you to interpret and analyse the data you have collected for your Outcome.

We will follow this by covering the formats, conventions and design principles relevant to your solution. We conclude the chapter with a discussion of ways to generate design ideas as well as design tools you can use to develop your data visualisation.

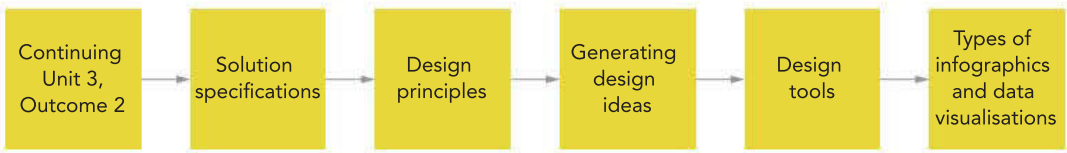


FIGURE 4.1 Chapter map

Solution specifications

The School-assessed Task (SAT) comprises two parts:

1. The solution for Unit 3, Outcome 2 is the information you create from the data you collect and analyse. This information determines support for your research question.
2. The solution for Unit 4, Outcome 1 is an infographic or dynamic data visualisation that communicates your findings about the research statement.

The following section discusses the specifications of the solution you will create for Unit 3, Outcome 2 to help you create a solution that meets the requirements and prepare you for gathering the appropriate data. When you create your solutions for the SAT, you will be using the problem-solving methodology (PSM) to guide their development. This discussion will include the *analysis* stage of the PSM, which involves designing the specifications of what is needed to produce a satisfactory solution.

- Solution requirements: What your information needs to do, the qualities it should have, and the data that is required to create the information.
- Constraints: The limits and restrictions under which the information must be produced.
- Scope: What the information must achieve, and what it is not required to achieve.

The following section discusses each of these specifications in turn.

Solution requirements

Any solution that you submit will have requirements that are functional and non-functional, as well as data requirements.

Functional requirements

Functional requirements describe the *tasks* that your solution should be able to perform. In the simplest terms, these are the things you want your solution to be able to do – the main reason you are creating it.

For example, an accounting program's functional requirements may specify that it must be able to:

- display summary totals and charts (**static visualisation**)
- access latest data from source (**dynamic visualisation**)
- change scale of charts (**interactive visualisation**)
- apply filters to select and display latest data (interactive and dynamic data visualisation).

The functional requirements are usually achieved in a specific and identifiable place in a solution, such as a particular formula, or control menu, dashboard, a piece of programming or a web page. The main functional requirement of your Outcome's information would be that it lets you reach a valid and substantiated conclusion as to whether your research question is supported or refuted.

Non-functional requirements

Non-functional requirements describe the *attributes* or *qualities* that your solution should have. Using the accounting program as an example again, its non-functional requirements may require it to be accurate, secure, fast and easy to use. A non-functional requirement will probably not be achieved in one specific place in a solution. It usually requires a combination of factors across an entire solution. For example, achieving ease of use in an accounting program may involve choice of simple menu and dashboard layout (for example, radio buttons and simplified charts) using design principles and making sure the interfaces are consistently user-friendly (for example, font size and **colour vision deficiency (CVD)** safe). The information you produce for your solution for this Outcome needs to be usable. This could mean information needs to be accurate, authentic, timely, relevant, complete, up-to-date (current), specific, unambiguous and unbiased.

Data requirements

The specifications you write for your solution must include a description of the data that you require to support your research statement. You must locate this data for your Outcome. You must find, select, reference, organise, process and interpret the data to produce information to enable you to state your findings.

For example, if you wanted to use the research question about Pavlov's theory from page 127, you would need the following data:

- the length of time that patients spend with animals
- levels of stress hormones in patients' blood
- subjective reports of stress.

9780170440875

If you wanted to use one of the vegan diet examples such as: 'Do increasing numbers of people in South Melbourne eat a vegan diet because they believe it has health benefits?', as a bare minimum, you would require data on:

- the number of people in South Melbourne who eat a vegan diet, to support the idea that numbers are increasing
- subjective reports of why people choose to eat vegan, to support or refute the idea that they are doing so because they believe it has health benefits.

You would also need to define the 'vegan diet' and 'health benefits' as part of your planning, and your subjects and respondents would respond to this information.

Interpreting your data

Interpreting your data may be a complex process depending on the research question you have chosen. You may find it difficult to be sure whether your research question has been supported or not.

When you are interpreting your data, consider the following questions:

- Have you been able to make a simple message about your findings from your research question?
- Can you make a statement that summarises your findings?
- Do your results meet your expectations?
- Do your results make sense?
- Could your results be interpreted differently?
- Is your supporting data of good quality?
- Is your supporting data current?

Resources to help analyse your data

You should also consider using tools for organisation and techniques for finding patterns, such as **mind mapping**. Patterns in data can be noticed more easily if you use tools such as **graphs**, or other visual techniques, such as **frequency distribution tables**, which will show the number of observations in each category.

Solution constraints

Review this section on constraints again when you are working on Unit 4, Outcome 1.

Constraints are limiting factors or conditions that need to be considered when you are designing a solution. A constraint will usually reduce your freedom of design choice. Constraints generally fall into five categories: economic, technical, social, legal and usability (Figure 4.2). (See Chapter 7, page 308–10 for further information on the APPs.)

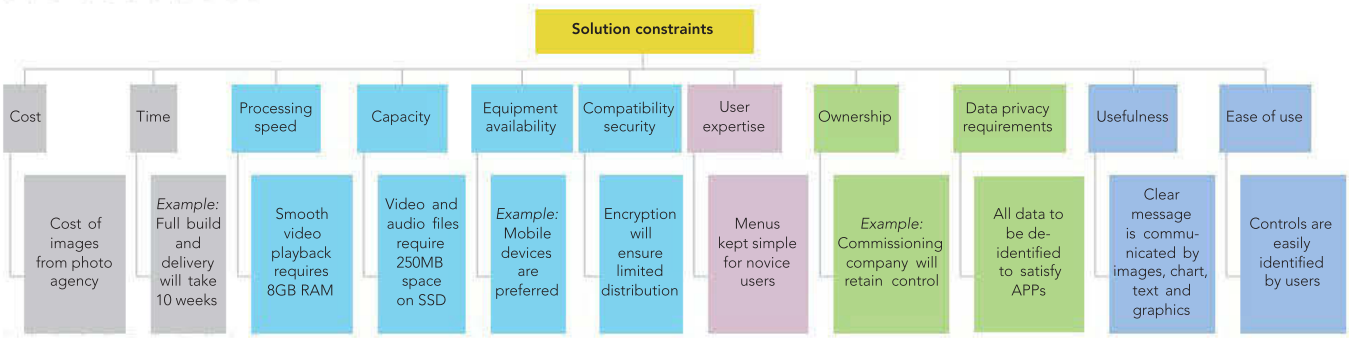


FIGURE 4.2 Constraints on a solution

Economic constraints

Some constraints are **economic constraints**. Perhaps your first consideration when designing your solution should be time. The deadline by when the user or client needs to have the solution operational will define the time available to design and develop the solution. The greater the timeframe, the more time there is to complete an in-depth analysis, detailed designs and develop advanced features of the solution. The shorter the timeframe, the faster that each stage in the problem-solving methodology needs to be completed. Be realistic in estimating your likely rate of progress. Are you assuming you will work more quickly than you can? Consider whether your solution relies on any other parties and whether they have any time constraints as well, such as whether you need to borrow any equipment and for how long. What are the costs of buying, maintaining and upgrading the equipment? What will it cost to pay consultants when you are designing your solution? You also need to factor in the costs of consumables. For example, if you need to do a lot of printing during the design of your solution, consumables include printer ink or toner. Both a lack of time or money may result in a re-evaluation of the user's requirements, or a re-evaluation of how the requirements can be achieved.

NelsonNet additional resource: Constraints on a solution table template

Technical constraints

Technical constraints relate to the hardware and software available for the project. Will the necessary equipment be available when you need it and what is it capable of achieving? Check that the equipment you want to purchase, rent or borrow is available in the early stages of planning. Capacity constraints also apply. Find out how much the system you are using can cope with in terms of disk space, bandwidth, CPU speed and memory. For example, large data sets may not fit into memory and still allow capacity for processing the data. This is a key difference between a database and a spreadsheet. For a spreadsheet, all the data is accessed at once so the data record must be held in computer RAM. In contrast, a database accesses one record at a time, and only the current record data is held in RAM. Data sets with millions of records can be manipulated in this way. The process takes time, however.

Most spreadsheet users expect immediate response from the application, usually due to the small number of records being manipulated. Several hundred thousand data items barely slow the modern notebook computer. Developers need to keep in mind that mobile users may not always have access to a high-speed network connection, so they need to ensure that any **animated visualisation** solution does not require a large amount of bandwidth to download and view.

4.1 THINK ABOUT DATA ANALYTICS

List some other technical constraints that developers of smartphone apps need to consider when developing a product.

Social, legal and usability constraints

Non-technical constraints relate to areas other than hardware and software. Usability and the user's level of expertise are examples. If a solution is being developed for users with little digital systems expertise, this may restrict some of the requirements that would involve complex manoeuvres to complete.

You may have **social constraints** that relate to level of expertise of users – some users may have limitations that affect how you design your solution, while others may have a concern about changes to existing solutions.

You must also meet any **legal constraints** by ensuring that your solution does not breach any copyright, data privacy or spam laws. Privacy laws may restrict features linked to displaying

9780170440875

THINK ABOUT DATA ANALYTICS 4.2

What problems do you think a clear scope of solution can avoid later in the project?

personal data in the solution, or to collecting data from the devices of someone using your solution. Copyright laws may restrict features that allow other users to upload content to the solution without the permission of the copyright holder. You must also not breach any of these laws yourself while developing your solution.

Finally, consider the **usability** aspect of your solution. Is your solution actually useful? Does it serve a purpose? If so, you must still ensure that your intended users will find it easy to use. It is unwise to build a solution for users who will find it unworkable. Any solution must plan to be user-friendly to as wide an audience as possible. This is addressed under 'Scope' below.

Scope

Identify what your solution will be and what you expect it to achieve. This is its scope. Remember that the solution for Unit 3, Outcome 2 is the information you produce from the data you have gathered to answer your research question.

Make sure that your research question clearly states what it includes, so that the information you produce does not go beyond the scope. A reasonable research question needs to be very specific. A research question that is too broad in scope makes it difficult to gather evidence to answer the question in the affirmative or negative.

The scope of a *solution* is largely defined by its functional and non-functional requirements, and may include lists of functions that are not required. For example, a solution may have far more data analysis options than the identified target audience may be interested in. A scope condition may limit the extent of the analysis. For example, only consider the last 10 years of computer records, rather than the last 50 years, due to relevance of results and skill levels of people involved.

Define the scope precisely so developers can allocate time and resources accurately and know when contractual obligations have been met.

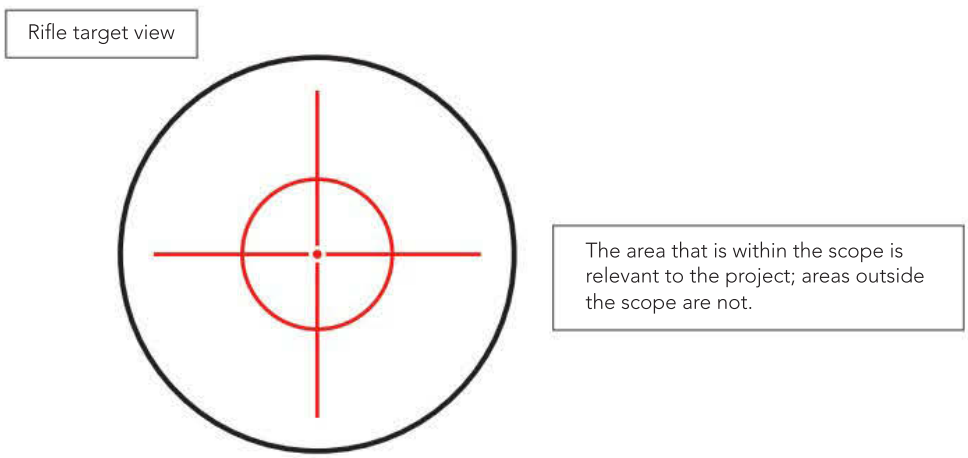


FIGURE 4.3 Scope of solution

Once you have settled on a research question, work out the scope of your solution. After you have started collecting both primary and secondary data for your Outcome, the next step is to design your solution. The following section of the chapter discusses design principles and criteria for evaluating efficiency and effectiveness of your infographic or dynamic data visualisation.

9780170440875

RESEARCH

Scope

The following sample research questions vary in scope, but all have potential problems. Discuss them in class. For each question, identify the problem, and then suggest either a replacement question or how the existing question could be tweaked to fix the problem.

1. Are Australian native animals dying out?
2. Are supermarket fruits and vegetables having their flavour bred out of them?
3. Is the internet changing everything?

CASE STUDY

Tattslotto

Tattslotto has been drawn many times since its debut in Australia in June 1972. The same number of balls each week have provided all the winning combinations. This investigation will explore every winning combination and attempt to identify patterns and make some predictions about the results of future Tattslotto draws.

- **Research question:** Are there patterns in the results of all the Tattslotto numbers ever drawn?
- **Functional requirements:** Only results from the official Victorian Tattslotto draws will be considered. Winning amounts will not be considered. All Saturday and mid-week draw results for every year will be included.
- **Non-functional requirements:** Data must be stored as efficiently as possible. Queries must be complete and accurate and contain relevant data.
- **Constraints:** The Tattslotto data must be manually recovered from a website. Data harvesting software, Outwit, will recover data in CSV format. Results are only available since draw 413, 6 July 1985.
- **Scope of the research question:** The data records will allow searches of combinations and summary records for each ball drawn. Data records will allow individual search and combination searches for each draw.



Tattersall's Sweeps Pty Ltd, https://thelott.com/tattslotto/results

FIGURE 4.4 The frequency each Tattslotto number has been drawn, and the number of draws since 6 July 1985

Design principles

Design principles are factors that enhance the appearance and functionality of solutions. Online solutions need to be easily understood and accessed with a minimum of time and effort. To communicate effectively, any infographic or dynamic data visualisation needs to be clear and functional. You need to ensure that facts are obvious and your message is unmistakable. Your solution must be carefully designed, taking into account an important set of design principles.

SCHOOL-ASSESSED TASK TRACKER

☑ Project plan	☑ Collect complex data sets	☑ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

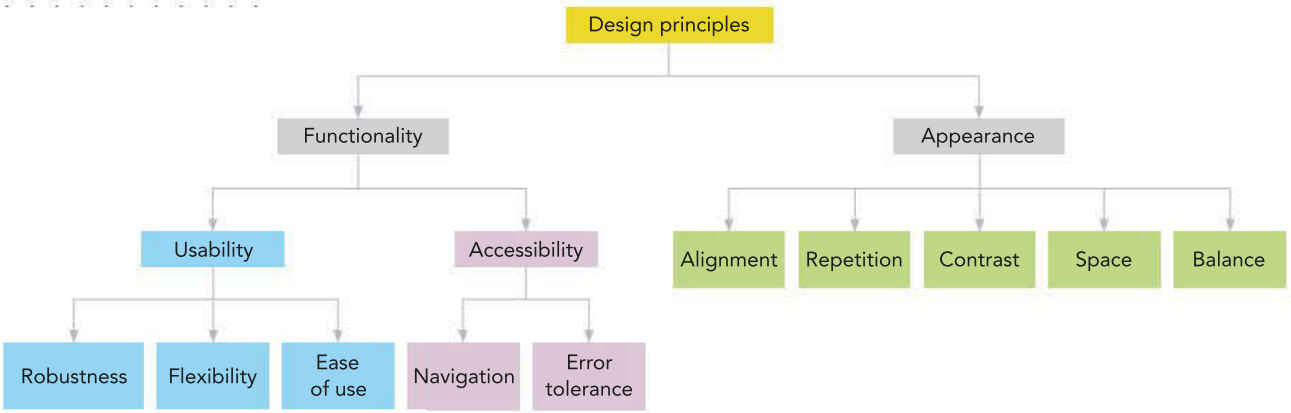


FIGURE 4.5 Design principles

If you sense that a page or screen is awkward to use, or looks odd or unattractive, but cannot exactly say *why*, you are probably responding intuitively to your solution's use of design principles.

In Unit 4, Outcome 1, you are required to create two or three design ideas from which one is chosen to be developed into a fully detailed design. A design idea is a rough, general 'back of an envelope' plan of a solution that has little detail. Techniques for generating design ideas are discussed later in this chapter (see page 176).

Functionality

Usability

The usability design principles are robustness, flexibility and ease of use (Table 4.1).

TABLE 4.1 Usability design principles

Principle	Description
Robustness	Your solution's ability to cope with errors during use. Robustness countermeasures include comprehensive data validation, preventing errors (such as by disabling certain buttons on an interface when they are irrelevant or harmful), and anticipating troublesome user actions (such as by choosing to print when no printer is connected, or saving to a device that is already full).
Flexibility	Your solution's ability to cope with multiple ways of performing tasks. For example, many websites were once designed with fixed-width dimensions, such as 800 px. You need to consider how the user will view your solution. If using a browser, a dashboard will need to re-size to allow for different sizes. For example, consider whether you will allow portrait or landscape orientation for mobile devices.
Ease of use (interactivity)	How user-friendly your solution is. Some of the user-friendliness aspects overlap slightly with the appearance design principles. However, you should also ask questions when creating, planning and testing your solution. • Is it easy for users to understand the design of your solution and perform interactive tasks? • Can interactive tasks be performed quickly? • Will users remember how your solution works when coming back after they have not used it for a long period of time?

9780170440875

Accessibility

The accessibility design principles are navigation and error tolerance (Table 4.2).

TABLE 4.2 Accessibility design principles

Principle	Description
Navigation	The clarity, simplicity and intuitiveness of your solution's navigation system. You should ensure your dynamic interactive data visualisation can be navigated by multiple browsers and a touchscreen (ensure the buttons are not too small and close together). High-quality user interfaces are transparent, meaning that your users would not really notice the actual interface because it is so easy to use that they interact with it intuitively. The required information would be found quickly enough that users focus on it rather than the way they found it. The interface is a navigation tool rather than an end in itself, so it should be unobtrusive but clear. For example, the labels on your navigation buttons should be short and clear; each page should include a link back to the start or homepage and this link should always be on the same place on every page. Use linked forward and backward arrows to indicate when text continues on another page, with 'Next', and 'Back' or 'Previous' to help orientate users. Do not hide the most important information under sub-menus or three screens down on the web page.
Error tolerance	Your solution's ability to help users avoid and correct mistakes using clear instructions, and its ability to prevent them making errors in the first place by avoiding allowing them to perform actions that could lead to errors. Grey out non-selectable options. Ask for confirmation of major actions. This is also connected to robustness.

Appearance

The screen layout of your dynamic data visualisation should not be unnecessarily elaborate or decorative, or contain superfluous animations. Too many buttons, bullets, icons, rulers and flashing graphics will confuse the eye and distract your audience.

Infographics can be printed as banners or posters or incorporated into on-screen displays. For on-screen infographics, long or wide screens that require your audience to scroll should be avoided when possible, particularly if scrolling is required both across and down the screen. This makes the information harder to view. Short screens of information, with links to other screens, are more easily viewed and therefore more effective.

The design principles contributing to appearance are alignment, space, contrast, repetition and balance.

Alignment

The alignment of text can be left, right, centre or fully aligned (justified). Generally, choose one alignment for each page and stick to it. Left-aligned text is easier to read than centred text for paragraphs because the text begins on the left-hand side every time. The text is in a straight line and readers can follow the text with their eyes starting from the left edge. Centred text makes the eye work harder to locate the start of each line. Unlike left alignment where there is a consistent straight edge for the eye to follow, there is no consistent focus point for eyes to return to once each line is read.

The human eye can detect when an object is only a single pixel out of place vertically or horizontally compared to its neighbours. Text, images, and columns should all be aligned precisely.

9780170440875

Sloppy alignment looks careless and unprofessional, and ruins the impression that items are visually related to each other.

The following text is centred:

> Many vegan-friendly restaurants and cafes have opened in the inner suburbs of Melbourne since 2017, such as Vegan Delight, which opened in August 2017 in Cape St, Fitzroy.

The following is the same text, shown right-justified:

> Many vegan-friendly restaurants and cafes have opened in the inner suburbs of Melbourne since 2017, such as Vegan Delight, which opened in August 2017 in Cape St, Fitzroy.

Full justification looks formal and is traditional in novels and textbooks such as this one, but in magazines and newspapers, it can lead to unattractive 'rivers' of white space travelling down narrow columns since text is stretched out to create a straight right margin.

Headings are often shown centred.

Repetition

Your audience will be reassured when repetition reinforces consistency in your solution. This is not repetition of content and words – rather, of design elements. Using the same logos, icons, typefaces, heading styles, colour scheme, margins, borders, menu positions and shortcut keys throughout your solution will help your audience to trust the predictability and consistency of your solution (Figure 4.7). This also ties into formats and conventions.

Essentially, your audience does not want to have to reorient and learn the system themselves on every screen they visit, and they do not want to see a different design on every page. This is especially the case when you want them to be able to focus on the content, not the design.

Contrast

Contrast refers to the visual difference in colour or tone between objects (both text and images). Greater contrast will make objects appear to stand out more from one another. If there is not enough contrast between two objects, they may appear to blend into each other, making it difficult for the user to see each of them clearly. Contrast between the background and text should make the information clearly visible and legible (Figure 4.6).

Take care when placing text over images. It may make the text unreadable. You should also avoid using certain colour combinations. Some combinations can be hard to read for everyone, and near impossible for the 1 in 12 males and 1 in 200 females who have a form of colour vision deficiency.

Colour vision deficiency, or CVD, affects approximately 5 to 8% of the male, and 0.5% of the female, Australian population. Alternative colour schemes and palettes can be discovered at ColorBrewer

ColorBrewer

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Pellentesque blandit nibh nisi, eget blandit sem tincidunt sed. Ut ac dolor at ipsum lobortis consequat. Vestibulum dignissim, eros quis fermentum pellentesque, odio ligula vehicula odio, a vestibulum nibh dolor non tortor. Phasellus vel libero vitae dui aliquam rutrum vel sed sapien. Fusce a diam porta, dictum ante vitae, vehicula purus. Suspendisse vehicula dapibus accumsan. Quisque at tellus nisi. Nulla dapibus ultrices ipsum eleifend dapibus. Nam facilisis pulvinar turpis eget lacinia. Fusce.	Lorem ipsum dolor sit amet, consectetur adipiscing elit. Pellentesque blandit nibh nisi, eget blandit sem tincidunt sed. Ut ac dolor at ipsum lobortis consequat. Vestibulum dignissim, eros quis fermentum pellentesque, odio ligula vehicula odio, a vestibulum nibh dolor non tortor. Phasellus vel libero vitae dui aliquam rutrum vel sed sapien. Fusce a diam porta, dictum ante vitae, vehicula purus. Suspendisse vehicula dapibus accumsan. Quisque at tellus nisi. Nulla dapibus ultrices ipsum eleifend dapibus. Nam facilisis pulvinar turpis eget lacinia. Fusce.

FIGURE 4.6 Avoid using light-coloured text on a white background, as shown on the left, because there is not enough contrast to make it readable.

9780170440875

Space

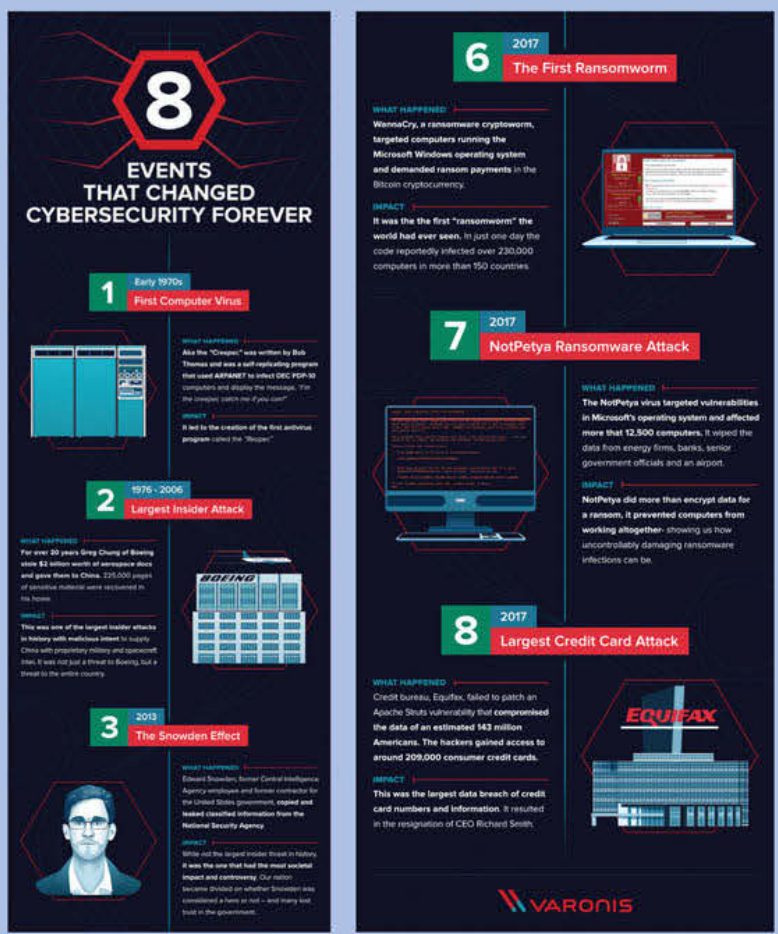
Space refers to the areas around and between objects – text and images (still and moving). If your solution is cluttered, it may be unpleasant to browse. You do need to include all of the information obtained for your Outcome on your solution, but you still need to space your objects so they can be individually distinguished and navigated through correctly.

Provide space between objects – columns of text, buttons on an interface, headings, and so on – so they are easy to perceive, but not overlapped and obscured. Each screen should not be so crowded with objects and features that the audience finds it difficult to see the information they need. You can use white space as a contrast to draw the user's eye. In graphics, animations and videos, levels of colour and contrast should make the information clear and attractive.

Repetition

RESEARCH

Annotate the two sections from the cyber security infographic in Figure 4.7 to show how repetition has been used to achieve consistency.



Used with permission of Varonis.

FIGURE 4.7 Two sections of a cyber security infographic

SCHOOL-ASSESSED TASK TRACKER

☑ Project plan	☑ Collect complex data sets	☑ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

A common convention is also to avoid yellow or other light colours for text on a white background because this can be difficult to read on a screen.

A large area of white space may be used to balance a section that contains an equally large area of text, because it will be of equal visual 'weight'. Working with space means also working with balance.

Balance

A solution with a balanced design is visually appealing. Solutions with unbalanced designs can lack the appropriate emphasis, look untidy and may discourage your intended audience from exploring them.

All elements of a layout have a visual weight. If the elements on either side or the top and bottom of the screen are of an equal weight, then visual balance is achieved.

There are two types of balance: symmetrical and asymmetrical.

With symmetrical balance, the visual elements on each side of an imaginary horizontal or vertical dividing line appear to be exactly the same in terms of visual weight, right down to the proportions and shading.

Asymmetrical balance occurs where visually matched weighting occurs through a combination of objects of differing sizes, shapes and colours. A large image slightly to the left can be balanced by an object far on the right. Humans enjoy symmetry in nature and music, and they crave it in design. You can use asymmetrical balance to draw your audience's attention to an exceptional item.

Generating design ideas

A **design idea** is a brief outline of a strategy for solving a problem. It lacks the detail and precision of a detailed design, but it points in the general direction of how a solution may be created.

Creative design techniques

There are several techniques for generating a range of creative and appropriate design ideas. The VCE Applied Computing Study Design does not name specific design techniques that you must know, but these are the most common techniques. All of them aim to find the most effective and efficient solution to an information problem. Your techniques should take into account the functional and non-functional requirements of your solution.



RESEARCH

Functional and non-functional requirements

1. Use the definitions of both functional and non-functional requirements as set out in this chapter, under the 'Solution requirements' section (page 167) as a guide to help you, identify:
   a. the functional requirements of your solution
   b. the non-functional requirements of your solution.
2. Justify your decision.
3. Discuss the functional and non-functional requirements of your solution in class with others (for example, your teacher and classmates).
4. Suggest how a design technique could take into account functional and non-functional requirements.

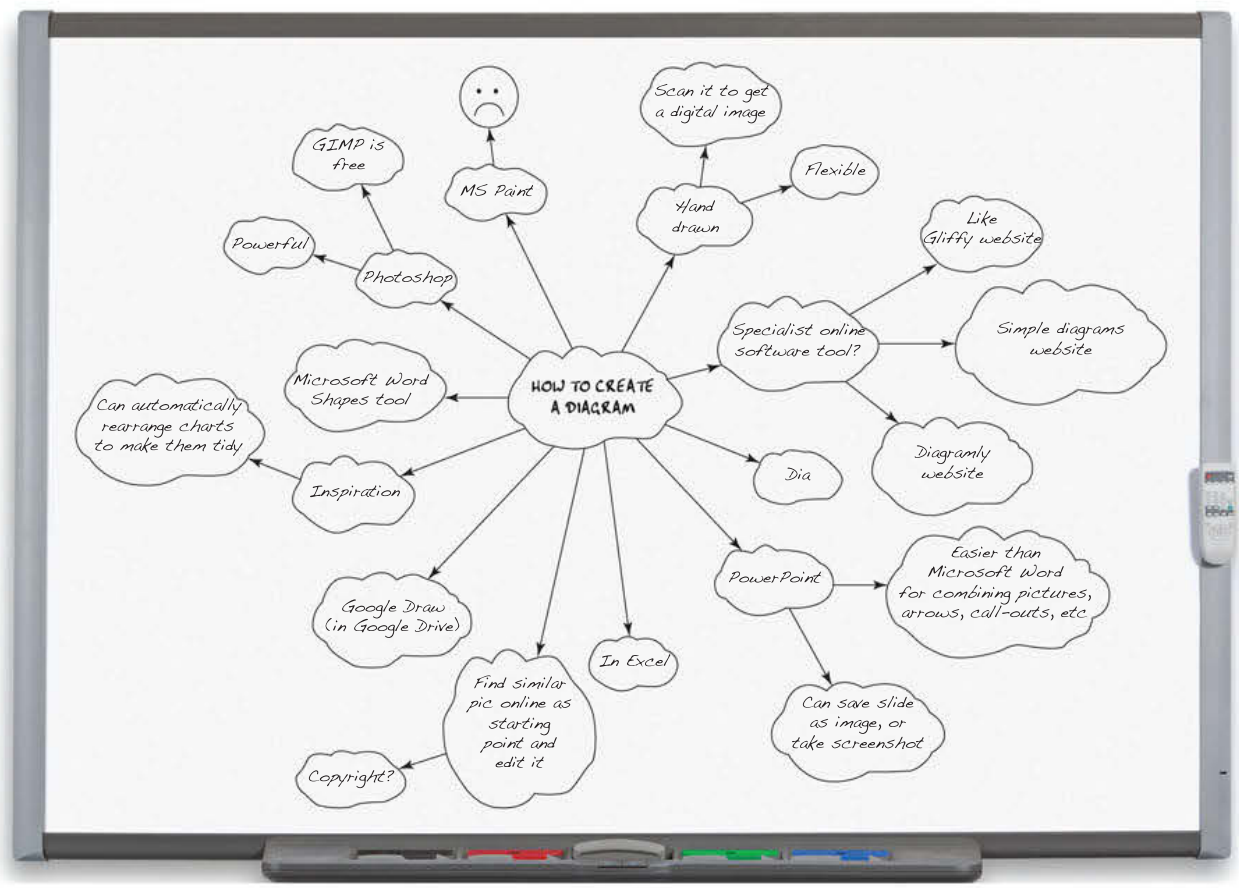
9780170440875

Brainstorming

Try not to hamper your imagination by rejecting ideas too soon. Brainstorming is a process where ideas are presented in a non-judgemental, spontaneous, unstructured and admittedly somewhat haphazard process.

Participants must have no fear of being judged, making mistakes or breaking rules. While some or even many ideas *do* turn out to be ridiculous, sometimes a half-baked, half-comical concept may turn out to be an unexpected work of creative genius, or it may stimulate a related idea that would be perfect. After all, these are only design idea suggestions, not the final design.

There are certain rules that you need to adhere to when you run a brainstorming session. The most important one is that no-one judges any contribution. No idea is criticised or rejected; every idea, no matter how outrageous or silly, goes onto a list of possible solutions. An idea that may seem slightly crazy at first can sometimes be workshopped into a great idea. In the 1970s, a brainstorming session came up with the idea of a pet rock. The idea was workshopped and before long you could buy not only a pet rock but a pet rock house and a training manual. Everyone just *had* to have a pet rock in the 1970s, and the idea made millions of dollars. It was the pre-technological version of the 1990s Tamagotchi (the handheld digital pet).



Shutterstock.com/Rtimages

FIGURE 4.8 Whiteboards are popular brainstorming aids because they are visible to all, and are simple and easy to edit.

Make sure that everyone listens to everyone else's ideas. Specify that only one person talks at a time and there is only one idea at a time. This not only ensures the shyest member of the group will contribute, but also makes it easier to record the ideas. Using these rules will help to elicit a number of workable ideas.

Brainstorming example: NASA faces the ongoing problem of lifting people and equipment into space. It is hugely difficult, expensive, loud and dangerous. How can it be improved?

Brainstorming for this project includes the following ideas.

- Helium balloons – float up, take off
- Fire rockets from the tops of mountains – reduce the distance to space
- A very, *very* tall ladder
- A giant catapult
- Antigravity capability
- A jet airliner to carry the rocket ship as high as it can, then the rocket takes off from there

While antigravity has no foundation in real science, some of the other design ideas could work, and would deserve more research. The team chuckled at the funny 'very, *very* tall ladder' idea until one person paused and said:

> Wait … I wonder if we could somehow get a super strong cable from the ground to low Earth orbit and anchor it in space, like a space elevator. You would ride up the cable to the end. The rocket can take off from there. You don't need all the fuel to achieve escape velocity … no need to launch from the ground.

'And … and re-entry,' said someone else. 'You could ride down the cable to get home. Simple. And low cost.'

From thinking what was whimsical, impromptu, unconventional and unconcerned with constraints comes a serious concept that has been further investigated by scientists at the Shizuoka University in Japan, with deployment of a prototype in October 2018. They aim to have a fully functional space elevator by 2050.

Brainstorming is helped by inviting people with different skills, experiences and areas of expertise into the team. Sometimes, a group of specialists struggling for a solution may be inspired by an idea from someone who is not constrained by their shared assumptions, preconceptions and modes of thought.

Consult end-users

Your solution, and all information solutions, will be used by real people. Thus, it makes sense to include real people in the design stage rather than wait for the *testing* and *evaluation* stages of the PSM to find out what they think of the solution. Manufacturers, political campaigns and film producers are known for their use of 'focus groups' of ordinary consumers whom they gather together and question about their likes, dislikes and reactions to design ideas.

A dedicated team of specialist designers may have their own ideas of what an end-user wants, but you should value primary evidence of your audience's needs and requirements.

Mind mapping

Mind mapping is ideal for complementing the process of brainstorming.

Mind mapping is a technique for generating and linking ideas. It is a creative and flexible tool that enables you to add, connect, organise and reorganise ideas. Mind-mapping software is generally flexible enough that you will not need to stop very often to learn how it works

while mapping; in other words, when mapping, your creative flow would not often be interrupted.

Unlike physical sheets of butcher paper or whiteboards (Figure 4.8), electronic mind maps can stretch endlessly in any direction (Figure 4.9), easily add or remove links between items, allow entire branches of thought to be moved to new locations, and you will not face the laborious task of copying out all of the scribbled ideas at the end of the brainstorming session (or you could take a photo!). The mind map can be saved for later development, printed, or transferred to a word processor for inclusion in a report.

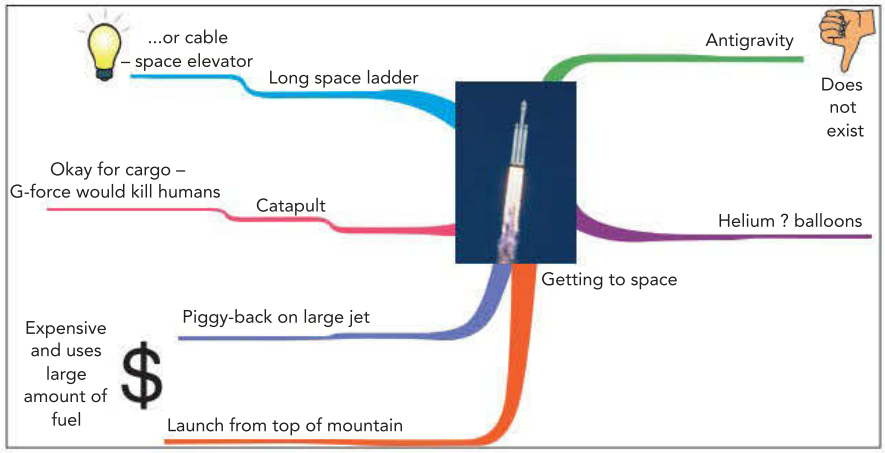


Photo: Alamy Stock Photo / NASA Photo

FIGURE 4.9 Mind mapping a project about how to get into space. This example was created using Inspiration software. Other software options include Lucidchart, Scapple and Bubbl.us.

Using the 'getting to space' problem, a mind map of the design process may look like Figure 4.9.

Lucidchart
Bubbl.us

Graphic organisers

Graphic organisers are visual methods of organising ideas. A popular graphic organiser is a PMI. A PMI (see Table 4.3) involves organising ideas into three columns: what has been successful (**P**lus), what was unsuccessful (**M**inus) and what needs more thought (**I**nteresting). You can use a PMI to reflect and evaluate, or to brainstorm new ideas.

TABLE 4.3 Example of a PMI

Using helium balloons to reach space		
P	**M**	**I**
Quiet	Crashes if gas leaks	Can balloon go high enough?
Relatively cheap	Slow to reach stratosphere	How much does helium cost?
Limited payload weight		

A KWHL is another popular technique of structuring thought, focusing on the scope of an investigation, and distilling the results of the research. KWHL provides a specific structure for thinking about a topic.

- **K** – What do I **k**now about the topic already?
- **W** – What do I **w**ant to know about the topic?
- **H** – **H**ow will I find out this information?
- **L** – What did I end up **l**earning?

A Venn diagram is a way to represent the similarities and differences in a set of concepts or objects graphically. The intersection of two or more circles provides an opportunity to list any shared features (Figure 4.10, page 180).

9780170440875

A spider diagram (Figure 4.11) is a powerful tool that can be used to create an overview of a central idea. The body of the spider is the central idea and the branching legs radiate out to related ideas and sub-ideas.

There are dozens of variations of such visual tools to help organise and clarify ideas. Others include character maps, concept webs, POOCH (**P**roblem, **O**ptions, **O**utcomes, **Ch**oice), ranking ladders (to prioritise or rank ideas, information or tasks), stair steps (to organise a process step-by-step), a chain of events, sequence charts (to put sequential factors in order), pie charts (to represent the relative sizes of components in a whole), bone charts, organising trees, and even Gantt charts for managing project timelines. Gantt charts were discussed in Chapter 3 with regards to project management.

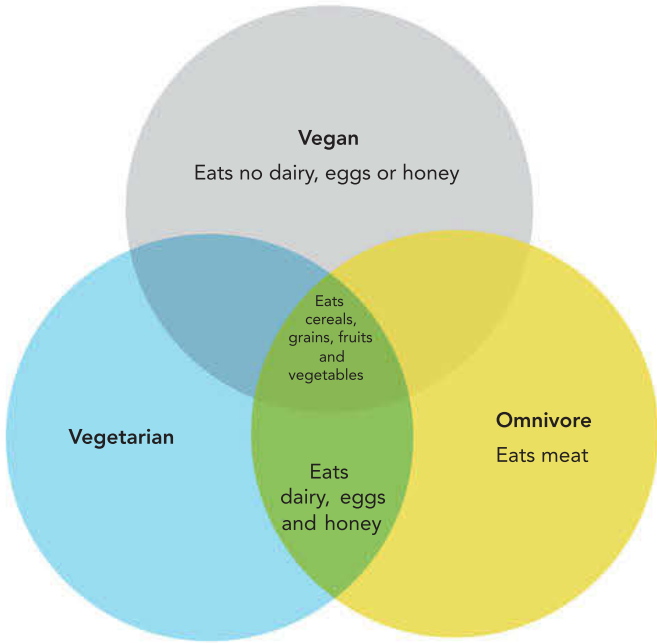


FIGURE 4.10 A vegan diet Venn diagram, showing its relationship to omnivorous and vegetarian diets

THINK ABOUT DATA ANALYTICS 4.3

Create a spider diagram similar to Figure 4.11, using drawing software. Your options are to choose from installing an application or an online drawing app.

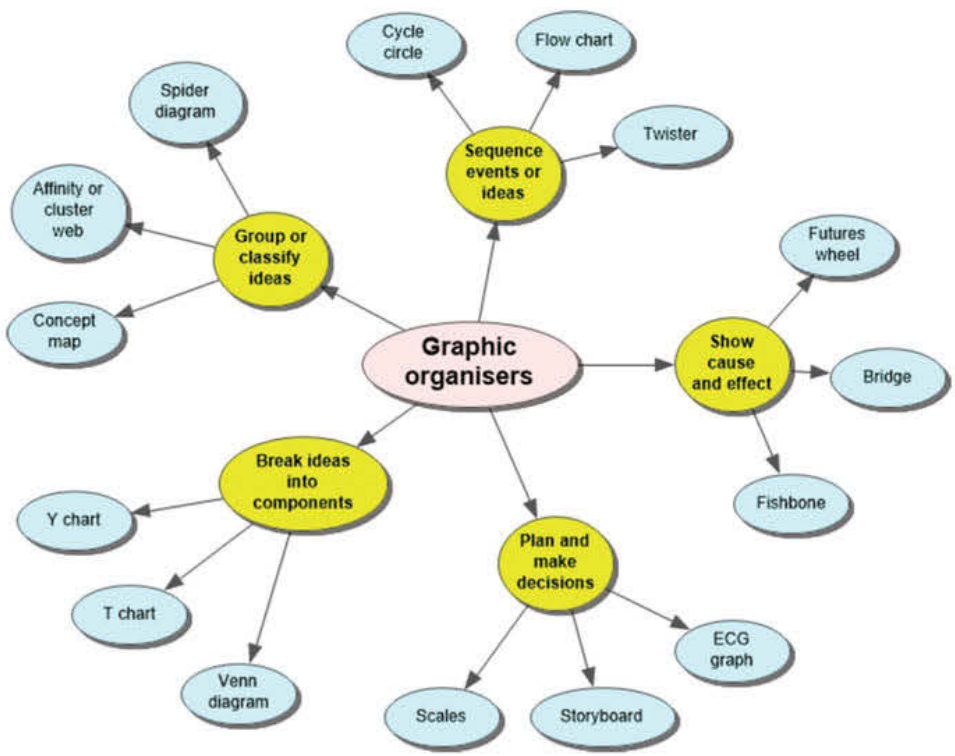


FIGURE 4.11 A spider diagram showing related concepts and sub-classes of concepts; this one was created with Inspiration software.

9780170440875

Attribute listing

Attribute listing helps break down preconceived notions about the nature of a product and helps to create fresh, new products.

Create a table with columns that make up a list of the attributes of a solution or strategy. In each column, fill in as many examples of that attribute as you can. Now mix and match one entry from each column to create the specifications for a brand new item. For example, you want to create a new game. You might create a table similar to Table 4.4.

TABLE 4.4 Game concepts

Genre	Age group	Platform	Gender target	Violence	Price range
Excitement	Toddlers	Phone	Male	None	Cheap
Problem-solving	Early primary	Tablet	Female	Comic, cartoon-like	Medium
Comedy	Late primary	PC		Realistic, mild	Expensive
Education	Young teen			Ultra-violence	
Strategy	Late teen				
Gambling	Young adult				
Role-playing	Middle aged				
Simulation	Elderly				

From this, you may stimulate thinking about a cheap game about gambling for young female teens on a tablet with realistic, mild violence, or you may pursue an expensive, non-violent, role-playing game for retirees on a PC. Obviously, not all combinations seem feasible, but an odd mixture might just give rise to a brand new concept, such as the untapped market for a comedy phone app for middle-aged women.

de Bono's Six Hats

Edward de Bono, renowned creative thinker and inventor of 'lateral thinking' devised a powerful problem-solving strategy known as 'Six Hats'. It is based on the observation that problem-solving people can think in different ways, and they can change those ways deliberately, just as if they were changing their hat for another. When a problem needs to be thoroughly investigated, people in the team are assigned to wear one of the six hats temporarily, each of which has a style of thinking that balances the thinking of a different hat-wearer (see Table 4.5, page 182).

Having representatives of each thinking type (colour of hat) ensures that a team is not neglecting certain ways of thinking. For example, a group of naturally creative, gentle and caring people may need the dispassionate thinking of logical people if they are to investigate all aspects of a problem thoroughly.

9780170440875

TABLE 4.5 de Bono's Six Hats

Hat colour	Focus	Questions that could be asked
Blue	Managing	What is the subject under discussion? What are our goals? How can we achieve those goals?
White	Information	What are the facts? What do we know so far? What data do we need?
Red	Emotions	How does that make you feel? What do you like/dislike about it?
Black	Caution and critical thinking	Why may that not work? What could go wrong? What are the problems and dangers?
Yellow	Optimism	What are the positives? What value will that bring us? What benefits will it have?
Green	Creativity	Where else could that lead us? What about if we ...?

Tips for creative thinking

Creative design can be learnt. You do not need to be born with the talent. There are techniques that anyone can use to improve their design creativity.

Substitute

Replace part of the problem with something else. For example, if you are producing hundreds of certificates, do not use mail merge to take data from a spreadsheet and insert it into a word processor – use a database instead.

Combine

Join unconnected things together, such as reducing the weight of camping supplies by combining a spoon and fork into a single utensil – the spork.

Adapt

Use an existing component in a different way, such as using presentation software (PowerPoint or Keynote) to create a poster. The first spreadsheet was created using the concept of paper-based accounting books.

Strip back to basics

Reduce the problem right back to its most basic parts and see what is left to address. For example, the tiny and cheap computer, the Raspberry Pi (Figure 4.12, page 183), is a stripped down Linux PC with minimal components. Inspecting the basics may reveal the nature of a problem more clearly.

Compare

Ask yourself, 'What other thing do I know that resembles *this* problem, and how does that other thing work?' For example, when sending a number of print jobs to a single printer, how can they be handled? Like a group of waiting people at a gate, you could organise them into a queue and process them in the order of their arrival.

9780170440875



Getty Images / AFP / MONA BOSHNAQ

FIGURE 4.12 A Raspberry Pi

Sleep on it

Creators often reach a point where they can make no further progress. Rather than dwelling on the same failed ideas, it is often better to let them go and think of something else. While the front of your brain is enjoying eating a banana or drinking a hot chocolate, or an episode of the latest reality cooking program, the back of your brain will be busily pulling ideas together to create a solution. The advice to authors when they suffer 'writer's block' is to go for a walk to 'clear their head', or to 'sleep on it'. After a fresh start, the ideas begin to flow freely again.

Research

Thomas Edison said: 'Through all the years of experimenting and research, I never once made a discovery. I started where the last person left off.' It is important to learn from your predecessors so you do not waste time 're-inventing the wheel'.

How have other people solved problems similar to the one you face? You are unlikely to be the first person in history to have faced a problem before. How have others coped? Their successes may lead you in the right direction, and their failures may prevent you wasting time. Care is needed when using your friend Google. Acknowledgements must be given and false information must be rejected.

Visualisation

Geniuses often make their thoughts visible because words cannot adequately convey the ideas they have. Einstein was famous for his non-verbal thought experiments. He visualised travel at the speed of light as travelling on a train. He said that written words and numbers did not play a significant role in his thinking process.

9780170440875

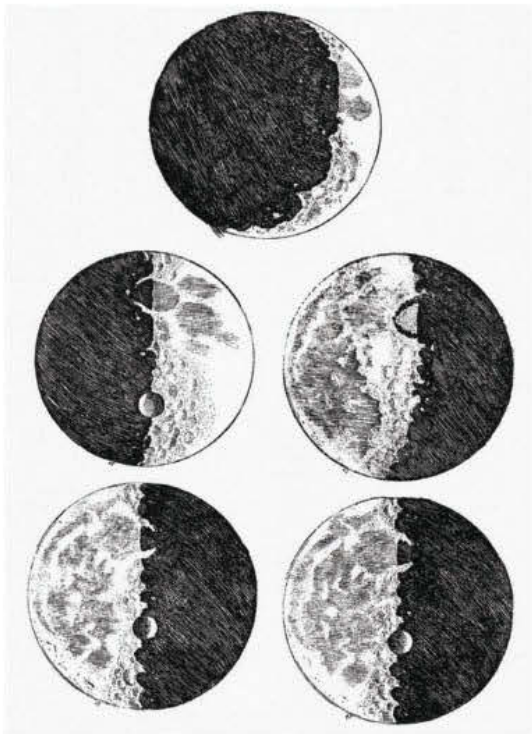
Leonardo da Vinci is renowned for his sketches of inventions. Galileo Galilei drew diagrams and maps of planetary orbits and phases of the Moon when others used mathematical formulas and words. Sigmund Freud, Alfred Hitchcock, Isaac Asimov, Beethoven and Mozart all reported the use of mental imagery in their creative processes. Dr Temple Grandin, famous for her work with livestock, said:

> I think in pictures. Words are like a second language to me … Language-based thinkers often find this phenomenon difficult to understand, but in my job as an equipment designer for the livestock industry, visual thinking is a tremendous advantage.

Temple Grandin, www.grandin.com/inc/visual.thinking.html

Locate the latest online drawing web apps. Consider whether you might use one of these to construct some of the drawings and graphics you will need for your infographic or data visualisation. Some issues to consider: how long will a trial version last? 15 days, 30 days, 3 months, 12 months? Will the graphic be able to be exported? What are the file formats of the exported file? Will a graphic created in the trial version be watermarked or not?

You may choose to use software simulations or models to help structure your thinking and construct knowledge.



Alamy/Stocktrek Images, Inc.

FIGURE 4.13 Galileo's drawings of phases of the Moon, based on observations through his telescope, 1610

Be observant and prepared

Many inventions arose from people seeing things that were similar to a problem for which a solution needed to be designed. Can a blockage in a canal be similar to solving a blockage in blood vessels? How can thousands of ants travel safely and quickly through a small gap, while a crowd of human spectators take nearly an hour to leave a football stadium? Discoveries are often serendipitous.

Sticky notes, potato chips, velcro, Teflon, cellophane, insulin, dynamite, stainless steel, super glue, cornflakes, vulcanised rubber and Play-Doh were all found by observant people after accidents or failed attempts to invent something else.

Research also suggests that creative people are typically hoarders – they keep lots of knick knacks, photos and articles around, and revisit these for stimulus at a later date.

Keep your eyes open and be receptive to connections between apparently dissimilar things. Revolutionary ideas often come from 'ridiculous' connections that no-one previously let themselves consider. Physicists argued whether light was a wave or a particle, until someone innocently (and correctly) proposed that it could be both.

Someone with a solid knowledge of a topic and an ongoing curiosity about new findings is receptive and can recognise the importance of an observation to an existing idea. The uncreative observer will either not notice the idea, or fail to see its relevance to a developing design.

Take risks, persist and be brave

A creative design idea often needs to take the risk of being dismissed, mocked or rejected. Many of the greatest breakthroughs were rejected at first and took a lot of time and effort to be proven right.

The germ theory, that diseases are caused by micro-organisms, was put forward by Louis Pasteur in the 1860s. It superseded the miasma theory, that a poisonous vapour in the air was the cause of disease. This theory had endured for several centuries. Pasteur's theory was initially mocked until further experimentation showed it to be most likely correct. In more recent times, Steve Wozniak combined the concepts of a typewriter, a calculator and a display. He was envisioning a whole new technological paradigm: the personal computer. His employer at the time, Hewlett-Packard, rejected Steve's concept five times. This led Wozniak to team up with Steve Jobs, which then led to the creation of Apple Computers. The idea of a tablet computing device had been tried by Apple and Microsoft and ended in failure. Steve Jobs tried again when the technology was mature, and the iPad was an instant success. James Dyson (of Dyson vacuum cleaner fame) is believed to have created 5000 prototypes of his vacuum cleaner over five years before he got it right. These examples show that it is persistence, not genius, that is probably the greatest contributor to success.

Thomas Edison, developer of the light bulb, phonograph and electric power, famously said:

> Genius is 1 per cent inspiration and 99 per cent perspiration.
>
> Many of life's failures are people who did not realise how close they were to success when they gave up.
>
> I have constructed 3000 different theories in connection with the electric light, each one of them reasonable and apparently likely to be true. Yet only in two cases did my experiments prove the truth of my theory.

Evaluating design ideas

When designing the solution to a problem, the first design idea you have is rarely the best one. A different strategy might be cheaper, easier, faster, more effective, or may better meet the client's demands. While one design idea may be attractive to the developer, the client may have non-technical constraints or priorities that will make one strategy more attractive than another. Providing a range of design ideas lets the clients choose the solutions that best suit them.

You may have used a design idea successfully in the past, but it may not be appropriate in the current circumstances. Although previously proven strategies can be useful, you need to be willing to think outside the box. Old strategies will not work for you in every situation – it is lazy and unimaginative to assume that they will.

A successful problem-solver will consider current functional and non-functional requirements and relevant constraints to develop an imaginative range of options from which the best design idea can be chosen and developed into a detailed design.

The criteria for choosing the best design idea may include:

- ease of use
- how long it will take to implement
- scalability (how easily the product can be increased in capacity)
- the degree to which it satisfies all requirements
- the degree to which it copes with constraints
- ease of implementation
- the amount of disruption likely to be caused to the organisation.

You will be using a version of this evaluation criteria as part of your assessment in Unit 4, Outcome 1. See page 248.

Evaluation questions from the functional and non-functional criteria include the following.

- Is the solution easy to use?
- Is the length of implementation time acceptable?
- Does the product satisfy requirements?
- Are the limiting constraints acceptable?

For a data visualisation and infographic some suitable criteria might include the following.

- Is the content well written and clear?
- Are the visuals appropriate for the target audience?
- Is the content informative?
- Is the content free of spelling or grammar errors, out-of-date or obviously inaccurate content?
- Are appropriate policies included and easy to understand (for example, security, privacy, copyright)?
- How is the user experience on different platforms?
- Is the content original or unusually engaging?

Some design decisions can be very difficult, and require careful balancing of competing needs – usually cost against quality. A design that is cheap and quick to produce may be barely competent, quickly wear out or be unpleasant to use. A superior design that would lead to a solution with a long life and happy users will probably take longer to produce and cost more.

Compare the likely differences in design philosophies and criteria between the pairs of objects shown in Figure 4.14 (page 187).

9780170440875

An F1 race car



Shutterstock.com/David Acosta Allely

A smart car



Shutterstock.com/Art Konovalov

A PC tower



Shutterstock.com/Den Rozhnovsky.

An all-in-one Mac



iStock.com/hocus-focus

A basic, inexpensive coffee table



Shutterstock.com/Mariyana M

An ornate, expensive antique table



Shutterstock.com/bergamont

FIGURE 4.14 Pairs of objects for comparison

Design tools

Input–process–output (IPO) charts

IPO charts help to design **algorithms** in spreadsheets, databases and programs, which can be used to devise formulas, scripts and program code. An IPO chart is created using the following steps.

1. Identify the information required, such as a person's age. List it in the *output* column.
2. Determine what input data is needed to calculate that output. For an age, that means a date of birth (DOB) and the current date. The data goes into the *input* column.

9780170440875

3 What processing needs to be done on the input to calculate the desired output? This algorithm goes in the *process* column, expressed in pseudocode: a mix of English and actual formula language to outline a way of calculating the information. Do not write full formulas; only the basic logic of the calculation strategy is required. Turn it into a proper formula later, during development.

Notice that in Table 4.6 there is one row for each output, and that output (from previous calculations) can be used as input in later calculations.

TABLE 4.6 Example IPO chart

Input (data)	Process (algorithm)	Output (information)
• DOB • Current date	Days between DOB and current date ÷ 365	Age in years
• Quantity • Cost per item	Quantity × Cost per item	Subtotal
• Is tax payable? • Tax rate %	Subtotal + (If tax is payable, then: Subtotal × Tax rate %)	Total cost
• Age in years • Total cost • Discount rate %	If age in years > = 60, Total cost – (Total cost × Discount rate %)	Senior citizen cost

Mock-ups or annotated diagrams

We discussed mock-ups and annotated diagrams in Chapter 1. Mock-ups or annotated diagrams show the intended appearance of printed output, on-screen information and interfaces.

A mock-up can be considered successful if you can give it to another person and they can create the interface without needing to ask you questions about it.

Mock-ups show features such as:

- positions and relative sizes of controls (buttons, scrollbars, status bars)
- positions, sizes, colours and styles of text (headings, labels, body text)
- menu positions and contents
- borders, frames, lines, shapes, images, decoration and colour schemes
- object alignments (vertical, horizontal, diagonal)
- contents of headers and footers.

Also consider that a flowchart may be the best way to plan the layout of your infographic or data visualisation.

Inspiration is a software tool that is useful for creating quick and easy design diagrams. In Microsoft Word, you can use Insert > Shapes. The drawing tools in PowerPoint or Keynote are another simple way of combining images, shapes and text to form a complex diagram. Either export the slide as an image (or Save As) or take a screenshot (PrintScreen key in Windows or Command+Shift+3 in macOS saves to desktop) and paste the image into a graphics editor for cropping and saving. (Crop the images using Command+Shift+4 in macOS: Use Command+Ctrl+Shift+4 to save to clipboard). The sequence of ideas can be arranged to offer the simplest, most easily understood message.

9780170440875

A mock-up can indicate location of text and images as well as necessary documentation including font style, colour and spacing.

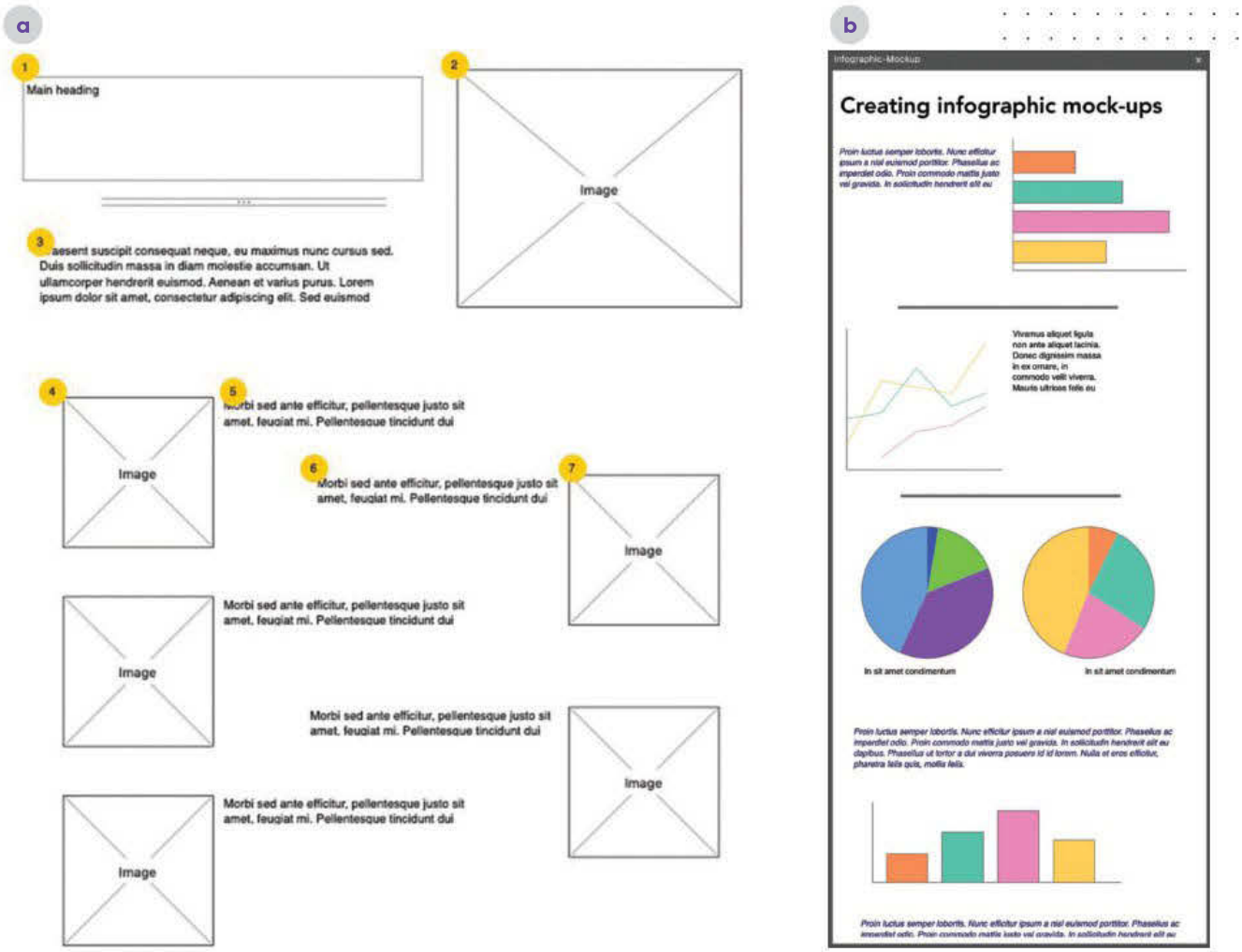


FIGURE 4.15 A typical mock-up for an infographic layout created in **a** Marvel, and **b** Moqups

Some suitable mock-up software tools (see weblinks) include the following.

- Canva – free for a single user and two designs
- Marvel – free for a single user and one project
- Balsamiq has a free educational licence available for teachers to arrange for their classes, and is renewable annually.
- Adobe XD – free (Minimum OS: Win10 64-bit or macOS 10.12 or later)
- Moqups – free (up to 400 objects) for interactive design with inclusion of placement for links and interactions
- Figma – free for a single user

Canva
Marvel
Balsamiq
Adobe XD
Moqups
Figma

SCHOOL-ASSESSED TASK TRACKER

☑ Project plan	☑ Collect complex data sets	☑ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

Types of infographics and data visualisations

While data visualisation (data vis or data viz) has many definitions, for our purpose, an effective rule for a 'good' data graphic is: 'Data visualisation means to tell a story with data and information'.

Before embarking on a design exercise, you need to be clear as to the purpose of the final product. There are many types of data visualisations, each with a specific purpose, depending on the desired outcome. Infographics, often in the form of a poster, can also be categorised for different purposes. You have already encountered some of these ideas in Chapter 1.

Data visualisation

The VCE Applied Computing Study Design has definitions of terms used in this study. Data visualisation is on page 8.

Data visualisation is a relatively new field of computing. A 'good' data graphic makes information accessible. Charts are no longer limited to static displays – they can be dynamic, interactive and animated. Every 'good' data graphic tells a story.

Figure 4.16 on the opposite page shows different types of charts used for different purposes, but these merely scratch the surface of types of data visualisations.

It is your choice how to best present the data you have selected to illustrate and support the statement of your findings. So, your challenge now becomes, 'How, and what, do you choose?'

Charts

Qualitative Chart Chooser 3.0

A **chart** (also known as a graph) is a method of displaying data visually, where the data set is represented as symbols. Many spreadsheet applications have chart (or graph) capabilities. Features of a chart can include a title, axis, scale or grid, data labels and a legend.

Charts are often used to visualise numerical data. There are a range of chart types, and each type can be used for different purposes. A *bar graph* can be used to compare different items, while a *pie chart* shows each data item as a proportion of the population. *Line graphs* are useful for showing the trend in a data item over time, and *histograms* are useful for grouping data then showing the frequency of each group. Refer back to pages 26–27 to review charts.

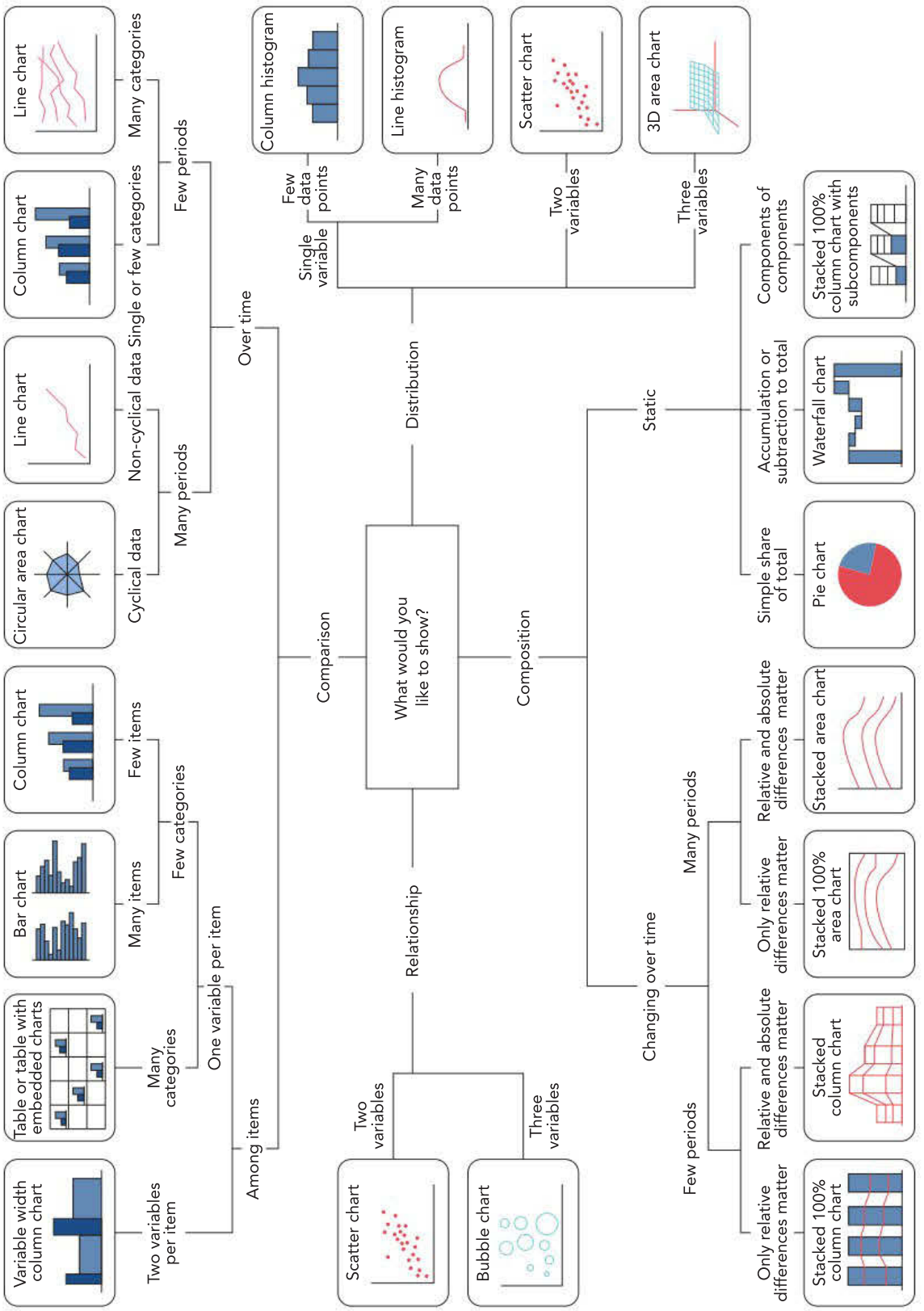
Map-based visualisation

While a picture may be worth a thousand words, you need to take care to ensure your intended audience can read, interpret and understand your relevant visual message.

A popular method to display geographical data is by using **map-based visualisations**. These types of visualisation are often called **geospatial visualisations**. Geospatial data is data that is related to the geographical location covered. Data could be related to population, roads, rivers, climate, mobile phone towers or any other characteristic of the area. Many geospatial visualisations are dynamic and allow the user to zoom in or out or navigate over an area.

Geospatial visualisations are becoming more popular because they are a very powerful tool that allows the data to be brought to life through visualisation. Since a range of data can be overlayed with a geographical location, the uses of these types of visualisations are enormous. Common uses have been for agricultural, environmental, mining and urban planning purposes, but the list is endless. An example of geospatial visualisation can be seen in Figure 4.17 on page 192.

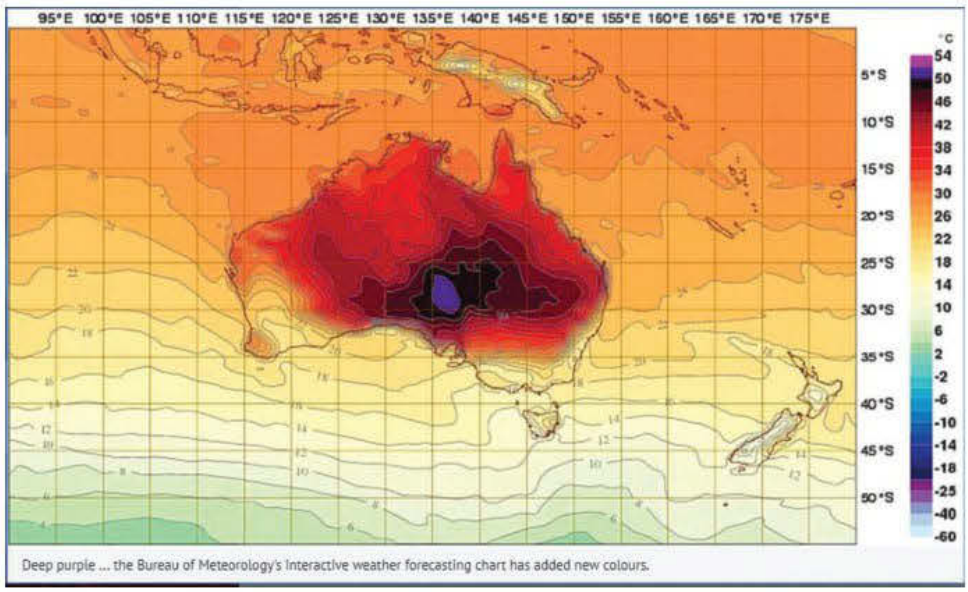
9780170440875



Advanced Presentations by Design: Creating Communication that Drives Action, 2nd Edition, Andrew Abela, May 2013, Pfeiffer

FIGURE 4.16 Some common types of data visualisations (not including network diagrams or maps)

9780170440875



Bureau of Meteorology, www.bom.gov.au

FIGURE 4.17 Geospatial visualisation (with the first use of mauve for above 50°C)

Network visualisation

Network visualisations show relationships between different data items and relationships between different data sets. Finding relationships within and between data sets has been an increasing area of interest in recent years as more data has become publicly available from both government and private organisations around the world.

A network visualisation might show the frequency with which individual players might pass the ball to each other in a football game or the number of people who travel on a public transport system each day. Network visualisations are also used to represent the layout of computer networks or public transport systems. Figure 4.18 is an example of a network visualisation showing a proposed plan for an NBN installation. The purpose of the illustration was to indicate how much closer, and shorter, a cable connecting a more easterly location than Sydney could increase internet access by a full second. This quicker access was estimated to be valued at over $1.0 billion AUD (2017).

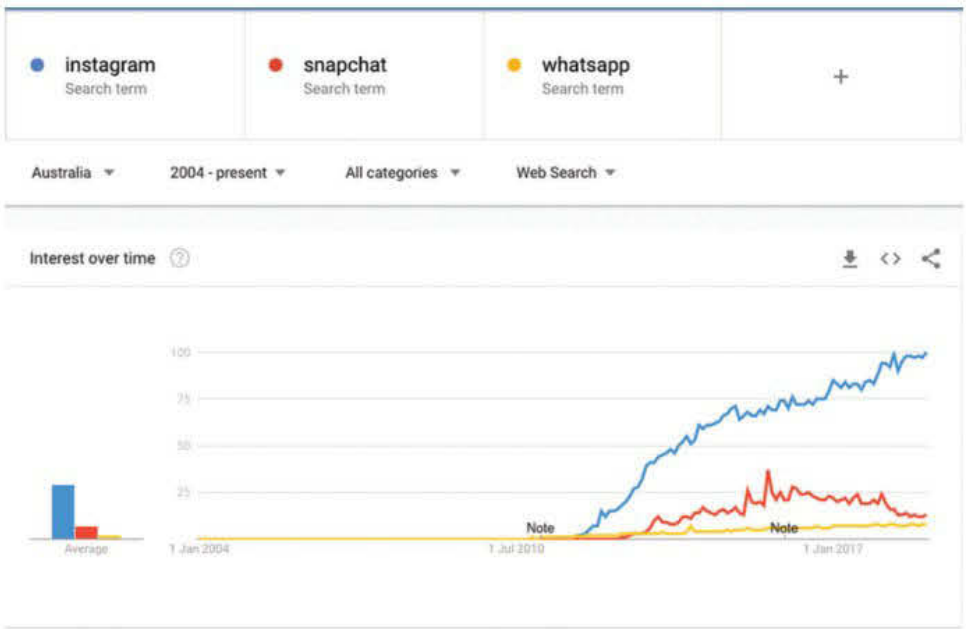


FIGURE 4.18 Network visualisation

Time series visualisation

Time visualisations represent a data item or data set over a period of time. Some time-based visualisations can show historical data, while others capture live data to provide real-time information. It is also possible to display the dimension of time by adding motion or animation to create a dynamic data representation.

The data could also be related to a timeline or time series. Timeline data may relate to individual items or events and show the order in which the items or events occurred over a time period, while time series data may relate to the same data item and show the variations or changes in the item over a time period.



Source: Google Trends

FIGURE 4.19 Time visualisation (with Google Trends on search terms between 2004–2018 for Instagram, Snapchat and WhatsApp)

Flow visualisation

Flow visualisations involve representing data to illustrate the flow pattern of a data item or items. This could be the pattern of customer movements through a supermarket or the series of pages a user would visit on a website to complete a transaction (user flow diagram).

Flow visualisations are also used for scientific purposes to visualise the flow patterns of objects that are normally invisible, including air and water. Figure 1.33 (page 35) represents the effect of an aircraft wing on the airflow passing the wing. The data for that visualisation was collected during testing using a wind tunnel and the data converted to a visualisation. Figure 4.20 (page 194) also depicts a flow visualisation.

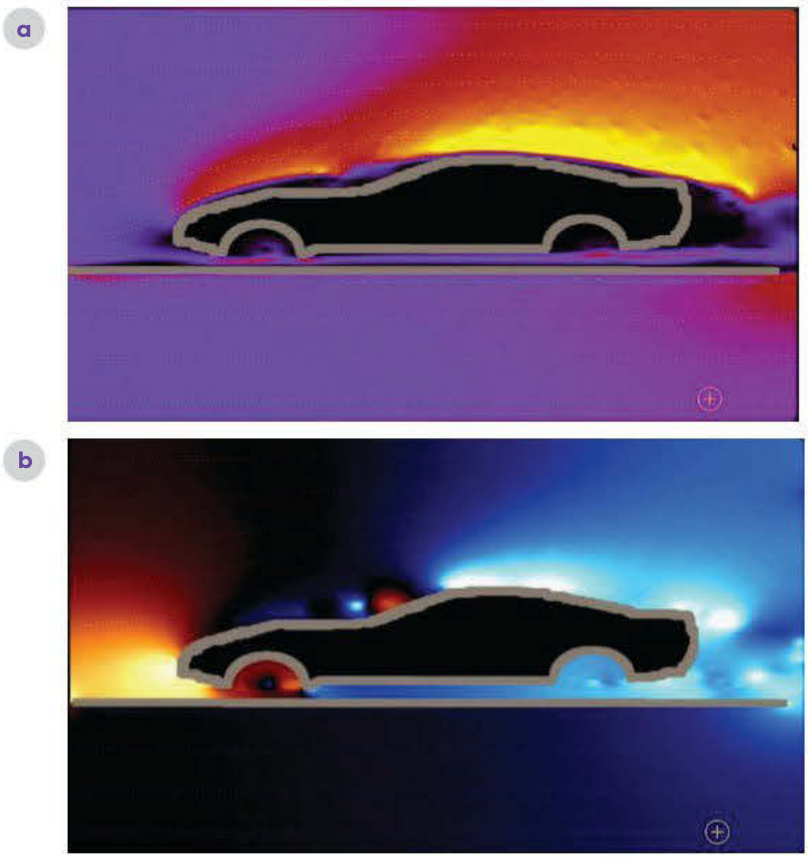
THINK ABOUT DATA ANALYTICS 4.4

- Using a search engine, find three examples of user-flow diagrams.
- Using Google Trends, investigate other social media search topic frequency changes.

SCHOOL-ASSESSED TASK TRACKER

☑ Project plan	☑ Collect complex data sets	☑ Analysis	☑ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875



Images created using the Wind Tunnel app © Algorizk

FIGURE 4.20
Flow visualisation of **a** air speed, and **b** air pressure

Matrix visualisation

Matrix visualisations can be used to show the composition of individual items in the sample size. Figure 1.26 (page 31) is an example of a matrix visualisation.

Infographics

A combination of the terms 'information' and 'graphic', an *infographic* seeks to convey understanding quickly and clearly to the observer. The infographic is usually a graphic representation of data, information and knowledge to be observed visually. Typical infographics take a poster **format**. There are several main types of infographic; however, the main elements used consistently include:

- graphs
- pictures
- diagrams
- narrative
- timelines.

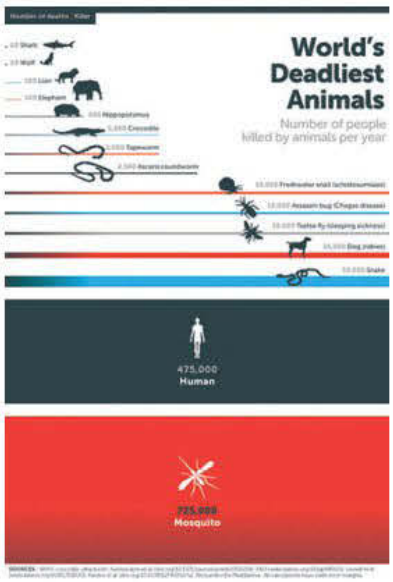
Infographics can incorporate multiple findings about a specific topic. Infographics in this form can communicate a more complete message about the data and information that is more relevant and engaging than traditional ways of communicating data and information (such as in a written report format).

9780170440875

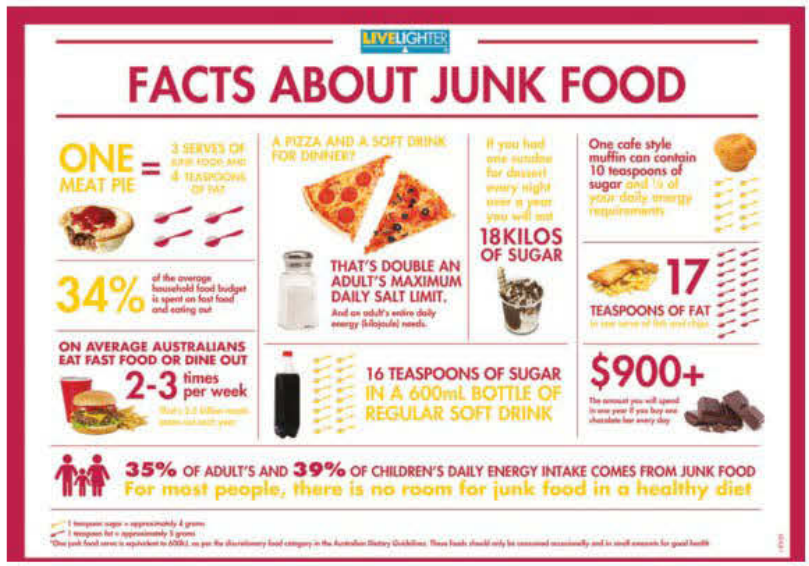
Statistical infographic

A **statistical infographic** emphasises the data such as the results from a survey. Readers should immediately be able to see the story that the data is telling.

a



b



World's deadliest animal graph courtesy of the Gates Notes.

LiveLighter® © State of Western Australia 2019, reproduced with permission.

FIGURE 4.21 Examples of statistical infographics

Informational infographic

In an **informational infographic**, information is appropriately displayed across a period of time. The time period may be for a project (weeks and months), an event (hours or days) or historical (hundreds, thousands or millions of years). Keep the timeframe consistent to avoid confronting the audience with the need to analyse and interpret a scale.



Illustrations and design by Scitech.

FIGURE 4.22 An informational infographic

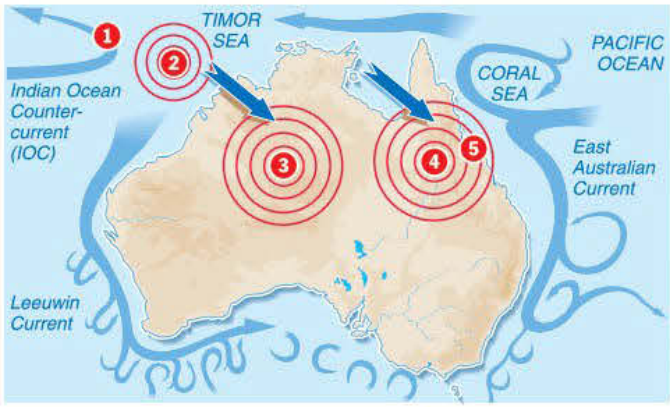
SCHOOL-ASSESSED TASK TRACKER

☑ Project plan	☑ Collect complex data sets	☑ Analysis	☑ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

Process infographic

Similar to the timeline infographic, a **process infographic** uses events rather than dates as the focus. Both employ the idea of the direction of flow with arrows and numbers to direct attention to the unfolding sequence.



MONSOON
Seasonal low pressure trough brings heavy rains from northwest

1 **October:** IOC blocks flow of water from Pacific to Indian Ocean

2 Relatively still water in Timor Sea heats up under tropical sun

3 **October-March:** Land heats up during summer. Hot air rises leaving lower air pressure behind. Low pressure heat trough pulls in moisture from Timor Sea

4 **December-January:** Low-pressure zone moves over Queensland, dragging in monsoon rains from north west

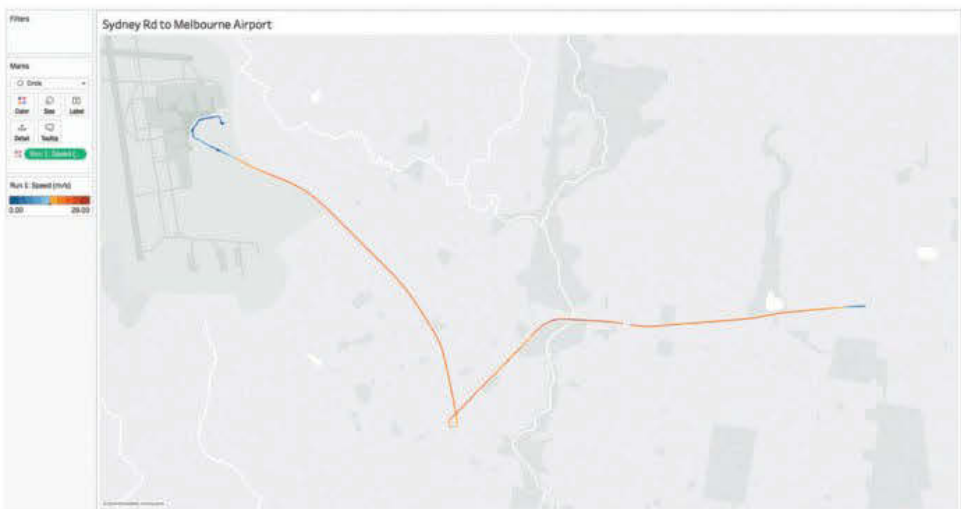
5 **2010: La Niña combines with monsoon rains**

Reproduced with permission of GRAPHIC NEWS.

FIGURE 4.23 A process infographic

Geographic infographic

The **geographic infographic** (also known as a geo-infographic) has the data and information included within a map. Google Maps, Google Earth and OpenStreetMap have greatly assisted with transferring coded data onto GPS coordinates.



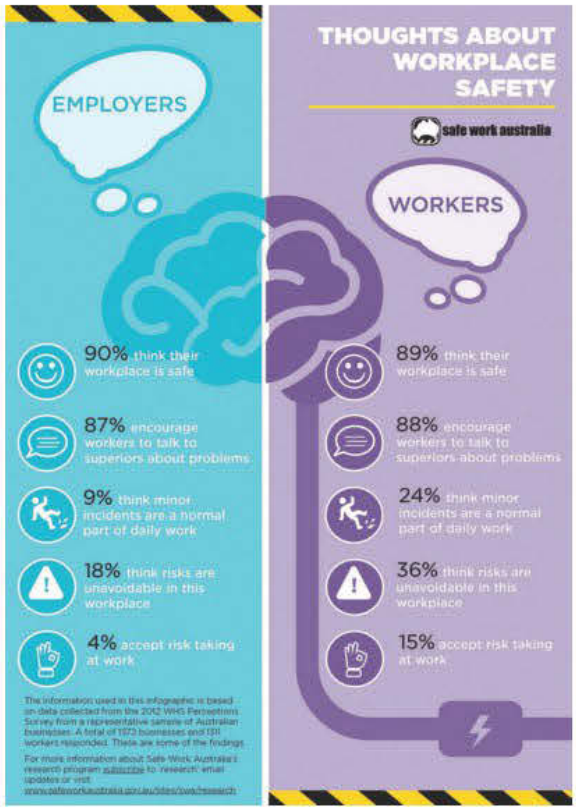
Created with Tableau.

FIGURE 4.24 This geo-infographic example is taken from a data logger in a suburb near Melbourne. Colour indicates speed – red = fast; blue = slow.

9780170440875

Comparison infographic

A **comparison infographic** places two data sets side by side so the viewer can compare and contrast the information.

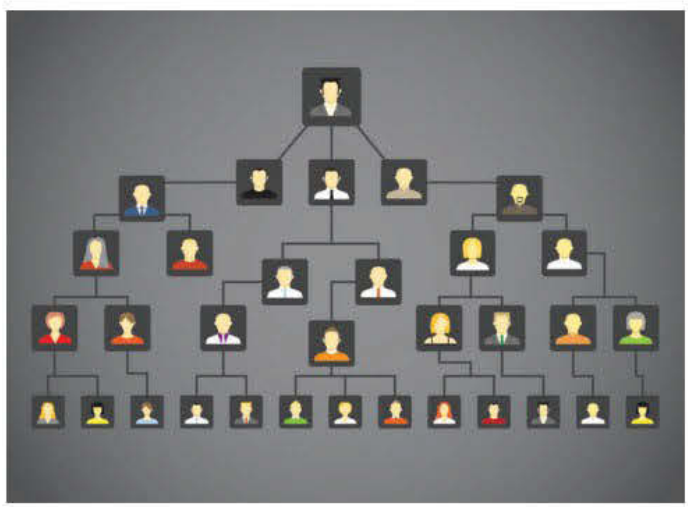


© Commonwealth of Australia 2018. Licensed under CC BY 4.0 licence.

FIGURE 4.25 A comparison infographic enables the viewer to compare and contrast.

Hierarchical infographic

A **hierarchical infographic** organises data so you can see the connection and level of the data. A good example is a family tree that shows connections between individuals and generations.



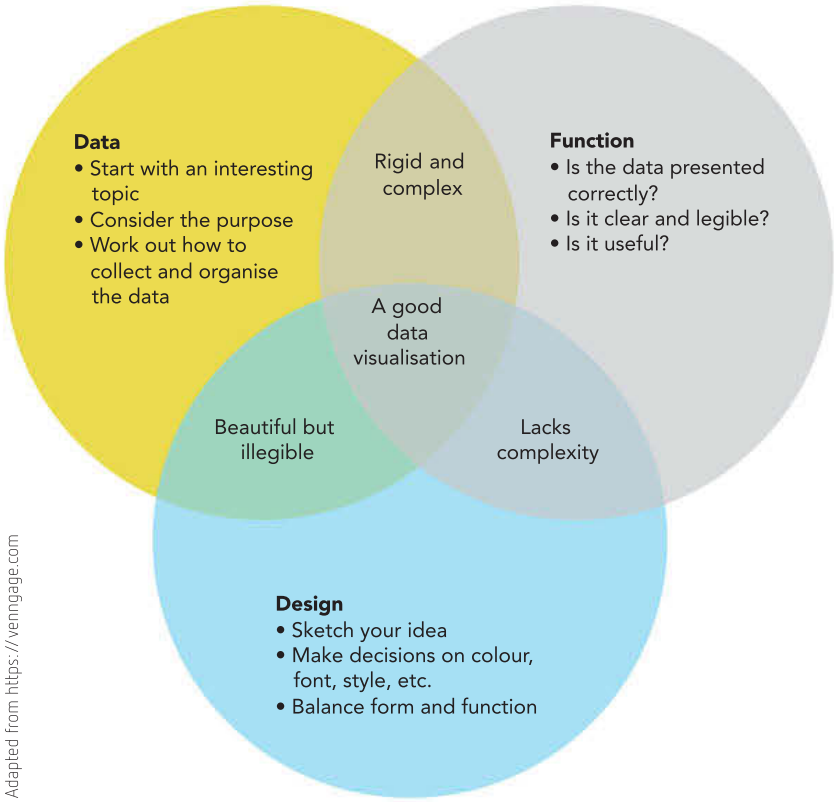
Shutterstock.com/tovovan

FIGURE 4.26 A family tree is a good example of a hierarchical infographic.

SCHOOL-ASSESSED TASK TRACKER

☑ Project plan	☑ Collect complex data sets	☑ Analysis	☑ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875



Adapted from https://venngage.com

FIGURE 4.27 A list infographic illustrating the relationship between the data visualisation elements: data, function and design

List infographic

A **list infographic** breaks up large lists of text by using bullet points, text and graphics to create a clearer understanding of the relationships between the data.

Choosing the best infographic to use

There are specific graphics that are suitable for certain types of information. Figures 4.16 to 4.28 indicate the range of options and how to decide. Once the category has been decided, there are visual considerations that may affect the final choice. For example, a graph can be a bar chart for non-continuous data (dates, categories) or a line chart for 'continuous' data. Data recorded every 30 seconds can be considered continuous (such as Figure 4.24).

Consider whether, when choosing a bar chart, would horizontal or vertical look better as an element of the infographic layout?

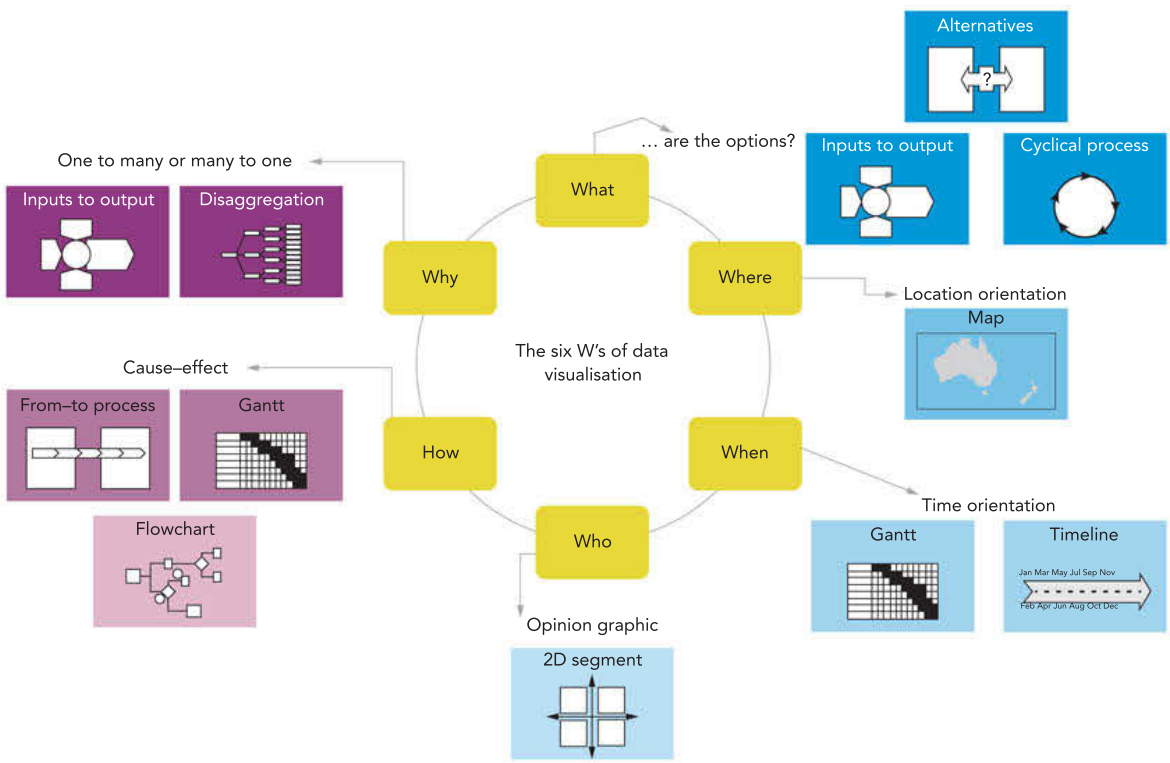


FIGURE 4.28 Use the six W's to help you choose which graphic to use.

Applications of graphics: Victorian train network

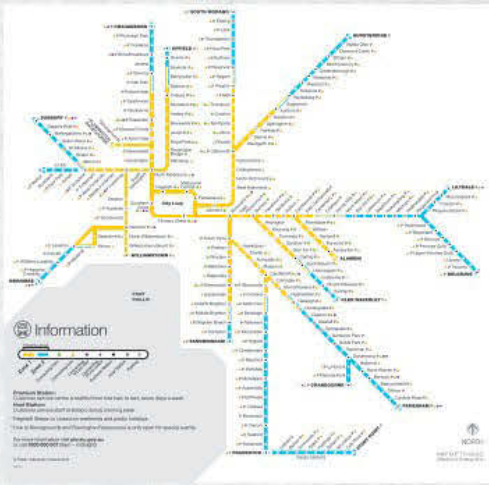
CASE STUDY

The suburban public transport network in Melbourne is used by hundreds of thousands of people every day for work, school and leisure. It is important that people are quickly able to view and understand information about the system that is important to them. They do not need to understand all the information. Figure 4.29b is the 2014 infographic that was used to show the lines and stations of the Melbourne metropolitan rail network. You will notice that the stylised map is not geographically accurate (Figure 4.29a).

a



b

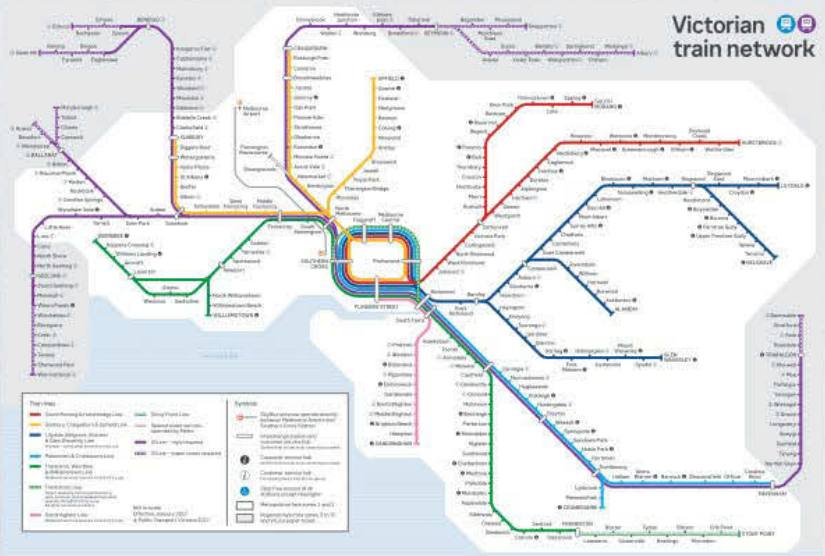


Alamy Stock Photo / Universal Images Group North America LLC

© Head, Transport for Victoria, all rights reserved. The train modal icons are trade marks of Head, Tfv and is used under licence by Cengage Learning Australia Pty Ltd.

FIGURE 4.29 a Satellite view of the geography of Melbourne; b the 2014 PTV rail network

One problem with the 2014 infographic was that it did not include country networks. The infographic was revised and expanded in 2017 to become the Victorian train network to show all lines and stations within Victoria.



© Head, Transport for Victoria, all rights reserved. The train modal icons are trade marks of Head, Tfv and is used under licence by Cengage Learning Australia Pty Ltd.

FIGURE 4.30 The 2017 PTV rail network map includes country trains as well. There is no connection to the physical geography of the areas represented in the graphic.

SCHOOL-ASSESSED TASK TRACKER

☑ Project plan	☑ Collect complex data sets	☑ Analysis	☑ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

4 CHAPTER SUMMARY

Essential terms

algorithm a description of a calculation strategy, such as a method of finding out whether a year is a leap year

animated visualisation a visualisation that shows the data story as a series of images

annotate to add comments to a document

chart a method of displaying data visually using symbols; sometimes known as a graph

colour vision deficiency (CVD) an altered colour perception; red–green is perhaps the best known, though there are many others

comparison infographic an infographic that compares and contrasts two options, typically left/right or top/bottom, which allows easier comparisons

data visualisation the presentation of data in a pictorial or graphical format

design idea creative exploration of concepts prior to selection

dynamic visualisation presentation that allows data and information to be updated, and the presentation adjusted to show the most recent findings

economic constraint a limitation imposed by time or finances

flow visualisation representing data in a manner that illustrate the flow pattern of a data item or items

format how something is displayed (for example, using a table or chart); arranging the look or presentation of an object

frequency distribution table a table that summarises values and frequency in two columns

functional requirement directly related to what the solution will do

geographic infographic an infographic using location to place data information; similar to a geospatial infographic, it visualises data onto countries rather than physical geography

geospatial visualisation where information is coded onto a map to illustrate the data (also known as map-based visualisation)

graph a visual representation of data showing the relationships between several elements

hierarchical infographic stacking items or subjects based on levels of importance, difficulty, income, and so on

informational infographic a list-based visual that is composed mostly of text

interactive visualisation allows the user to change settings and control the presentation

legal constraint limitations and restrictions imposed on a solution by laws and regulations

list infographic an infographic that breaks up large lists of text with bullet points, text and graphics to depict relationships between the data

map-based visualisation using location to place data and information to compare regional and global information

matrix visualisation a diagram that shows the composition of individual items in the sample size; in this regard, they can be considered similar to pie charts

mind mapping visually organising information into a diagram that shows links between the information

network visualisation depicting relationships between different data items and relationships between different data sets

non-functional requirement other requirements that the user or client would like the solution to have but that do not affect what the solution does

non-technical constraint a limitation that relates to areas other than hardware and software such as usability and users level of expertise

9780170440875

process infographic a flowchart that shows sequence in a decision-making process

social constraint a limitation imposed by social considerations

static visualisation a visualisation that does not change after the initial analysis and development of the presentation

statistical infographic a visual that combines images with statistics to improve visual appeal

technical constraint a limitation imposed by hardware or software on the solution

time visualisation a visual that represents a data item or data set over a period of time

usability a measure of how useful and usable a solution is

Important facts

1. **Dynamic interactive data visualisation** manually or automatically updates data and has user options for selecting prepared information to explore a presented message. 'Dynamic' allows for updating the data, while 'interactive' allows user control of the view.
2. An **infographic (information + graphic)** is a graphic visual representation of data, information and knowledge, often as a poster, that conveys understanding quickly and clearly. An infographic is often a static diagram suitable for printing as a poster or banner, or used as an interactive graphic on a web page.
3. A successful **solution** must convey its message clearly and be easy to understand.
4. Data visualisations make **data patterns and relationships** clearer than lists of numbers.
5. Choose **types of charts** with care so they visualise data clearly and accurately.
6. The main elements used for **infographics** include graphs, pictures, diagrams, narrative and timelines.
7. A **detailed design** is produced by the design idea that best solves the problem.
8. **Design ideas** may come from brainstorming, using outside consultants, talking to end-users, mind mapping, graphic organisers, attribute listing, de Bono's Six Hats, research and looking at parts of the problem in different ways.
9. **Persistence** and being observant increase the chances of devising the optimal design idea.
10. **Design principles** affect the functionality and appearance of solutions.
11. **Evaluation criteria** are rules set out during design that include effectiveness and efficiency criteria; they are based on the solution's requirements that were defined during analysis.
12. Each **evaluation criterion** will be expected to have a corresponding method.

9780170440875

TEST YOUR KNOWLEDGE

Review quiz

Solution specifications

1 List at least three functional requirements for an infographic or data visualisation.
2 List at least three non-functional requirements for an infographic or data visualisation.
3 List at least three technical constraints for an infographic or data visualisation.
4 List at least three types of non-technical constraints for an infographic or data visualisation.

Design principles

5 Suggest two ways to make a data visualisation dynamic.
6 List three ways in which an infographic may be changed to improve its accessibility.
7 Choose the three factors that contribute most to a usable interface. Justify your choices.
8 Explain the difference between formats and conventions.

Generating design ideas

9 What techniques could you use to generate an infographic design for a data set about a national sport?
10 How would you evaluate an infographic? List five criteria to choose your 'best' infographic.

Design tools

11 List three types of software that could design an infographic or data visualisation for your Data Analytics SAT. Choose one and explain why it is the best for your circumstances.
12 What type of design elements can be used in an infographic?
13 Why are mock-ups necessary when designing a solution? What detailed specifications could be included?
14 Where do the evaluation criteria come from? How can you ensure you have a process so you can choose between alternate designs?

Types of infographics and data visualisations

15 What questions need to be asked before beginning an infographic?
16 How does an infographic clearly convey its message?
17 Why are there so many different types of infographic?
18 What is the difference between a data visualisation and a dynamic data visualisation?
19 Why are there so many types of data visualisation?
20 How would you choose the 'best' type of chart for a data visualisation?
21 Why might an infographic be chosen instead of a data visualisation for a solution?

9780170440875

APPLY YOUR KNOWLEDGE



This will continue your work on the research question/topic you investigated in the Chapter 3 'Apply your knowledge' section.

1. Look back on the research topic you investigated in Chapter 3. Justify your choice of research question/topic.
2. Identify the independent and dependent variables for your research question.
3. Describe any constraints that affected your data acquisition, including availability of data, time and resources.
4. Explain the processes you used to locate, select and acquire qualitative and quantitative data.
5. Comment on the integrity of the data you acquired, including possible faults such as inaccuracies, biases and incompleteness.
6. Explain the software functions and techniques you used to process the data to generate support for your research question.
7. Summarise the functional and non-functional requirements of the information you generated from the data you acquired.
8. Complete the Gantt chart covering all of the tasks required for the 'Apply your knowledge' activities in Chapter 3 and Chapter 4. The Gantt chart should include task and task sequencing, milestones, dependencies, time allocation and resources.
9. Justify the file-naming conventions you used for all documents you created for the 'Apply your knowledge' activities in chapters 3 and 4.
10. Explain what led you to discover patterns and relationships in the data that supported any conclusions you were able to form about your research topic.
11. Summarise the hardware and software you used to input, store, communicate and output data and information for your research topic.
12. Explain what physical and software controls you took to protect the integrity of stored and communicated data and information.

SCHOOL-ASSESSED TASK TRACKER

☑ Project plan	☑ Collect complex data sets	☑ Analysis	☑ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

PREPARING FOR Unit 3

OUTCOME 2

Use a range of appropriate techniques and processes to acquire, prepare, manipulate and interpret complex data to answer a research question, and formulate a project plan to manage progress.

To achieve this Outcome, you will draw on key knowledge and key skills initially outlined in Unit 3, Area of Study 2. This Outcome begins the School-assessed Task (SAT) that will be concluded in Unit 4, Outcome 1.

Outcome milestones

1. After reading the detailed instructions provided by your teacher, identify a research topic or question and begin the Gantt chart that covers the whole SAT project.
2. Locate and document resources, process data and complete analysis to determine whether your research question is answered. Continue to update your Gantt chart.
3. Design a presentation of the findings as an infographic or dynamic data visualisation. Continue to document evidence for satisfying SAT criteria.
4. Finalise your Gantt chart.

Steps to follow

1. Read the instructions provided by your teacher. **Note:** No research topic or data sets will be provided.
2. The start and end dates of the SAT will be provided by your teacher. **Note:** Your teacher may determine when certain parts of the Outcome need to be completed, such as submission of the research topic and question. Schools will be provided with a form associated with authenticity. Work needs to be submitted and signed by you and verified by your teacher. These dates can be supplied by your teacher and included in your Gantt chart.
3. Begin your Gantt chart to plan and monitor your progress during the Outcome. As the project progresses, regularly update the chart to be a continuously accurate record. Be sure to include:
   a. project milestones within the start and end dates provided by your teacher, which also includes development tasks during Unit 4, Outcome 1
   b. identification of tasks
   c. sequencing of tasks, including the identification of concurrent tasks
   d. time allocations for tasks
   e. allocation of resources.
4. Generate a reasonable research question, keeping in mind that relevant data must be available to provide evidence to answer the question.
5. Begin your reference list to acknowledge intellectual property. You are not required to obtain copyright permissions, but you must acknowledge all primary and secondary data sources using the APA method.

9780170440875

6 Gather relevant, valid and reliable complex data from more than one source. Keep your reference list up-to-date as you search for data.

7 Organise and prepare the data for manipulation with an appropriate software tool, such as a spreadsheet or database.

8 Manipulate the data to produce accurate, relevant and meaningful information that will allow you to make statements about your research findings.

9 Interpret the information and conclude if the research question has been answered.

10 Write a short analysis report that sets out:
   - a a statement of your research topic and question
   - b the conclusion you have drawn from the information you generated from the data
   - c an outline of the findings supporting your conclusion.

11 Prepare for submission of your collection of data sets, and the information derived from them. Printing large data sets may prove to be unwieldy, slow, expensive and environmentally unfriendly. Consult your teacher about options for data file formats, packaging and exchange.

12 Explain the specifications used for creating the information. Include:
   - a functional and non-functional requirements
   - b constraints
   - c scope.

13 Finalise your reference list so that all data sources are acknowledged.

14 Show evidence of your data validation processes and techniques.

15 Show evidence of your data manipulation processes and techniques.

16 Explain the methods you used to secure the data and information you stored and communicated.

17 Show evidence of a project plan (Gantt chart) indicating times, resources and tasks up to the milestone that concludes Unit 3, Outcome 2.

Documents required for assessment

1 A short analysis report that sets out the definitions for the requirements, constraints and scope of your chosen solution, which can be either an infographic or a dynamic data visualisation.

2 A collection of complex data sets, and information derived from them, that allows findings to be made about the research topic question.

3 A folio of alternative design ideas with detailed design specifications for the preferred design.

4 A project plan (Gantt chart) indicating times, resources and tasks.

SCHOOL-ASSESSED TASK TRACKER

☑ Project plan	☑ Collect complex data sets	☑ Analysis	☑ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

Assessment

A more detailed set of assessment criteria will be available on the VCAA website under the Data Analytics study page. This will include mandated criteria and a marking rubric. There will be specific requirements about the complexity of the data and any software requirements. The SAT (comprising Unit 3, Outcome 2 and Unit 4, Outcome 1) will contribute 30% towards your study score.

9780170440875

Unit 4

INTRODUCTION

In Unit 3, Outcome 2, you collected data sets, in particular selecting, referencing and organising data for manipulation in order to prepare to determine the findings to a research topic or question.

In Unit 4, Area of Study 1, you will use the development and evaluation stages of the problem-solving methodology to complete the second part of the SAT. This will involve the manipulation and interpretation of data to determine the findings and the creation of an infographic or dynamic data visualisation to present the findings of the research topic or question identified in Unit 3, Area of Study 2.

You will especially focus on good design (chosen from several alternative design ideas) and communicating your message clearly to the intended audience. You will also evaluate how effectively your visualisation solution achieves its goals. Throughout the creation of your solution, you will continue to modify and evaluate the project management plan you began in Unit 3, Outcome 2.

In Unit 4, Area of Study 2, you will consider cybersecurity requirements for digital networks and the importance of data and information for organisations.

Area of Study 1 – Data analytics: Development and evaluation

OUTCOME 1 In this Outcome, you will develop and evaluate an infographic or dynamic data visualisation that presents findings to a research topic or question. You will also assess the effectiveness of the project plan in monitoring progress.

Area of Study 2 – Cybersecurity: Data and information security

OUTCOME 2 In this Outcome, you will investigate the current data and information security strategies of an organisation, examine threats to security and the consequences of not protecting data and information. As part of this, you must recommend strategies to improve current practices.



Getty Images / oxygen

CHAPTER

5 Development and evaluation

KEY KNOWLEDGE

After completing this chapter, you will be able to demonstrate knowledge of:

Digital systems

- procedures and techniques for handling and managing files, including archiving, backing up, disposing of files and security
- the functional capabilities of software to create infographics and dynamic data visualisations

Approaches to problem solving

- characteristics of information for educating targeted audiences, including age appropriateness, commonality of language, culture inclusiveness and gender
- characteristics of efficient and effective infographics and dynamic data visualisations
- functions, techniques and procedures for efficiently and effectively manipulating data using software tools
- techniques for creating infographics and dynamic data visualisations
- techniques for validating and verifying data
- techniques for testing that solutions perform as intended
- techniques for recording the progress of projects, including adjustments to tasks and timeframes, annotations and logs
- strategies for evaluating the effectiveness of infographics and dynamic data visualisations solutions and assessing project plans.

Reproduced from the VCE Applied Computing Study Design (2020–2023) © VCAA; used with permission.

FOR THE STUDENT

This chapter discusses the key knowledge required to complete Unit 4, Outcome 1. The second part of the SAT covers project management, evaluation techniques, data manipulation, how to write well for a diverse audience and how to develop an infographic or dynamic data visualisation.

FOR THE TEACHER

This chapter is based on Unit 4, Area of Study 1, and provides the key knowledge required to complete Unit 4, Outcome 1. The second part of the SAT continues the PSM stages with development and evaluation. The infographic or dynamic data visualisation will be developed by the completion of this chapter. Students will also complete an evaluation of their solution and the project management process.



Shutterstock.com/seamuss

Data visualisations

> The result of a process of using software tools to select and access data from large repositories to present the data as a graphic representation usually in the form of charts, histograms, graphs, maps, network diagrams and spatial relationships diagrams. Data visualisations help to identify patterns and relationships in large amounts of data. Data visualisation tools allow graphic representations to be static or dynamic and can incorporate virtual reality and augmented reality.

Reproduced from the VCE Applied Computing Study Design (2020–2023) © VCAA; used with permission.

Any graphic that displays and explains information, whether in data or text, can be described as a **data visualisation**. Data visualisations can take many forms; however, the elements used to convey the message can include any of the following.

- Graphics
- Images
- Video
- Audio
- Animations
- Data
- Text
- Charts

Charts can be constructed in many styles, and new ones are regularly developed that have not been seen previously as animation, and presentation technologies are enabling greater functionality.

Different types of charts and their purposes were listed in Chapter 4, page 191, Figure 4.16. These include the following.

- Scatter plot to illustrate relationship between variables
- Line chart to show the relationship between two continuous variables
- Bar chart/column chart for discrete and category data
- Pie chart for up to six categories of the same variable
- Pictograph using an image or graphic to illustrate data in columns/rows

Another less common chart is the rose (or coxcomb) chart (see Figure 5.1), which is a combination of stacked bar and pie chart used to convey proportion and amount in one image.

The software chosen to create your charts will be capable of reading, processing and displaying your data set. Some data in raw form needs to be cleaned or manipulated for the chart to be displayed (see Chapter 2). For example, some elements may have text instead of numbers recorded, or have an empty field. The chart can be created once all the data has been verified and cleansed, ready for processing. Microsoft Excel, Tableau Public and Wolfram Mathematica are readily available software that can provide tools to construct many types of charts.

9780170440875

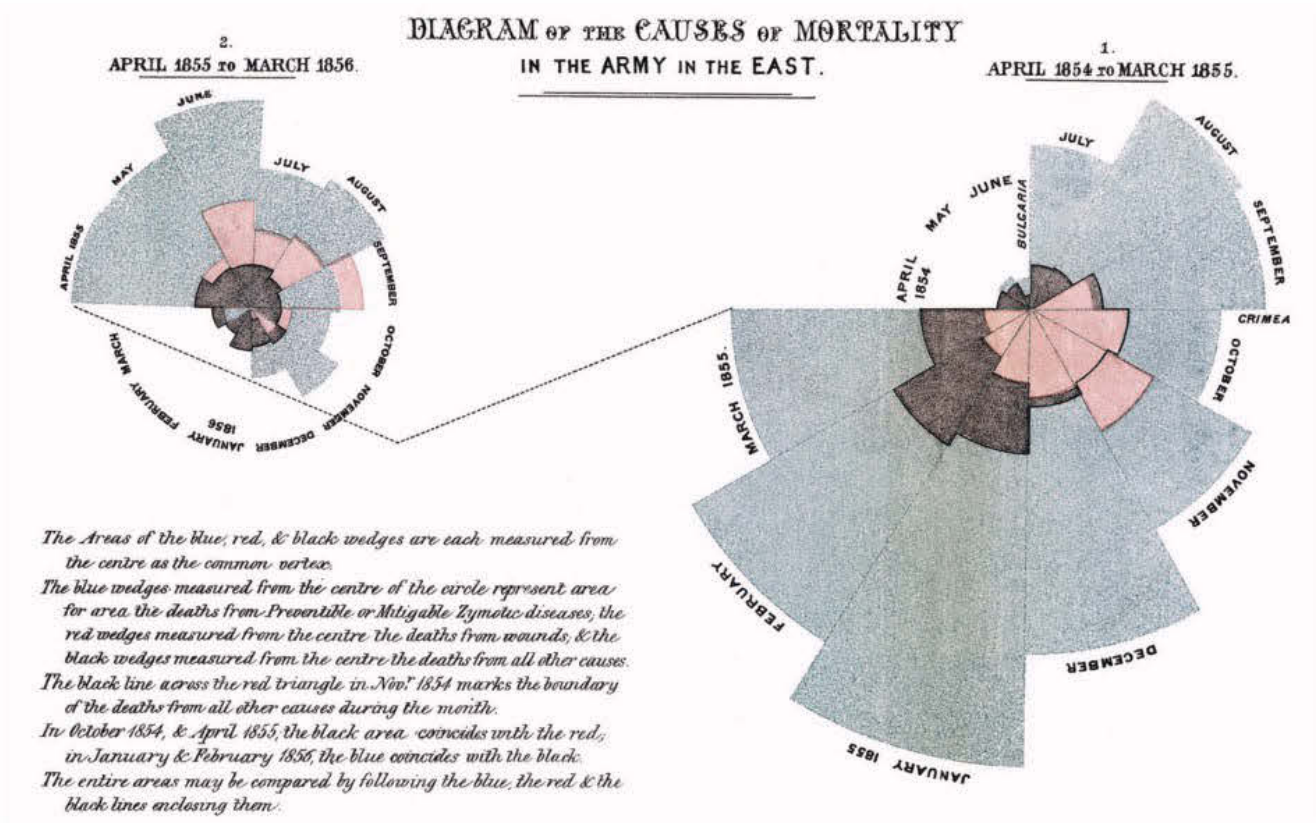


FIGURE 5.1 Florence Nightingale provides persuasive rose chart analysis arguing disease caused more deaths than wounds in the Crimean War 1854– 1856.

CASE STUDY

Chart with a story

Widely regarded as the best graph ever drawn, Charles Minard's visual representation of Napoleon's march to Moscow gives the user information on several levels at once: geographical, climatic and numerical.

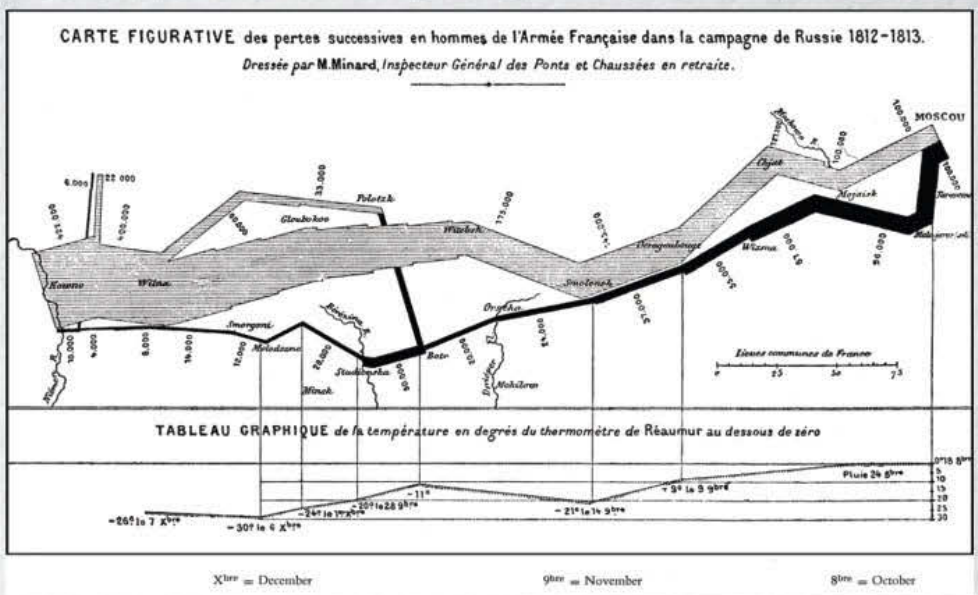


FIGURE 5.2 Napoleon's march to Moscow

The top line represents the troops moving towards Moscow in 1812–1813. The figures on the side of the line show the dramatic fall in troop numbers, and the corresponding thinning of the line shows the fall visually. It can be seen that Napoleon lost large numbers of troops each time he crossed a river. The line at the bottom of the graph shows the temperature. As the temperature dropped, still more troops were lost. The black line shows the troops retreating. One can only speculate about how different history might have been if Napoleon had had Minard in his midst during the troop campaign.

What is an infographic?

Graphical representations of complex data or information. They rely upon visual elements to quickly and clearly communicate patterns or trends in data or information. These include complementary colour schemes, easy-to-read fonts, graphs, simple charts and statistics.

Reproduced from the VCE Applied Computing Study Design (2020–2023) © VCAA; used with permission.

Infographics present complex data and information visually. Visual information can be quickly and clearly understood, with higher retention than other forms. The infographic could be as simple as a poster with fixed graphics, images and text. An interactive infographic can respond to user input and change the displayed information. This may be pre-set with dropdown menus or by allowing the user to input data or queries within a range of values.

THINK ABOUT DATA ANALYTICS 5.1

While referring to Figure 5.3, research how the updated 'Internet Minute' infographic may have changed.

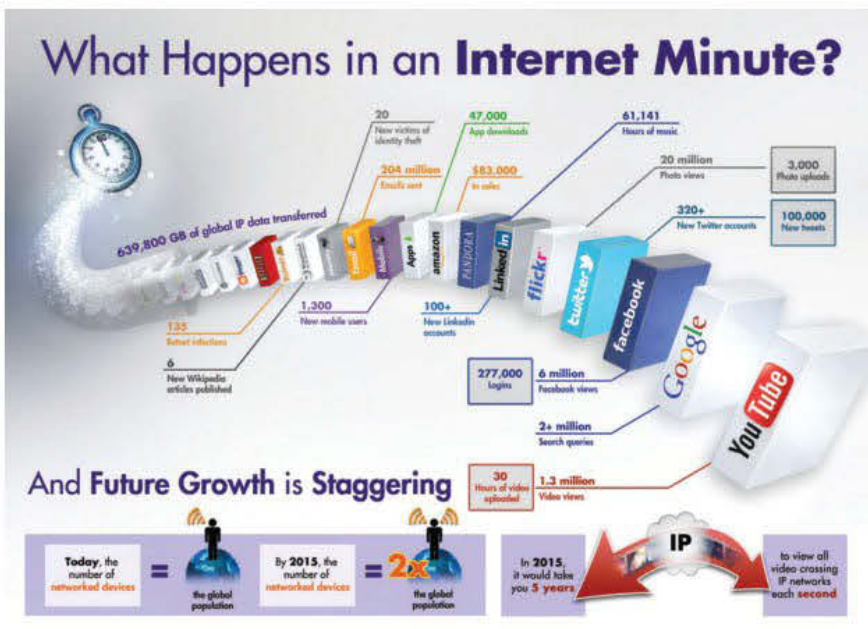


FIGURE 5.3 A creatively drawn infographic using colour, shapes, icons, lines and arrows, graphics and illustrations, with minimal text. This graphic was drawn in 2014.

IntelFreePress.Licensed under CC BY 2.0 licence.



www.domo.com. Used with permission of Domo.

FIGURE 5.4 More information does not mean easier to understand.

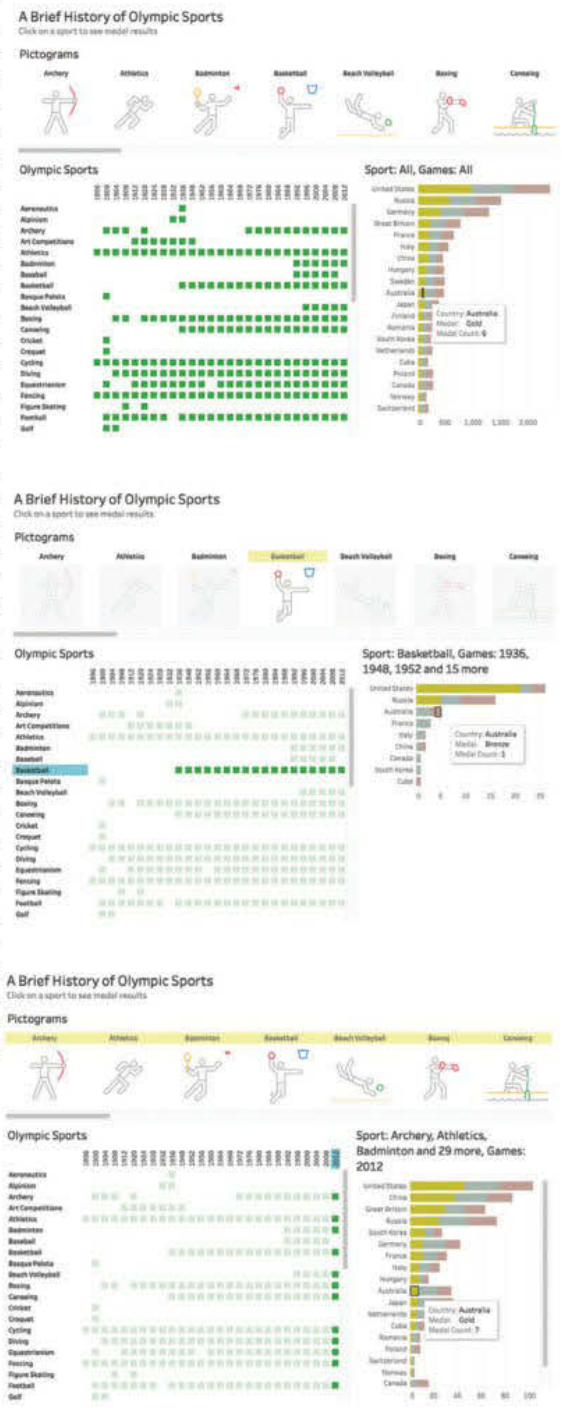
What is a dynamic data visualisation?

A **dynamic data visualisation** is where the appearance of a given infographic can be changed. This can be by the user selecting alternate views, which may have been previously processed, or by updating the data (or being continually updated) so the latest information can be included in the presentation.

SCHOOL-ASSESSED TASK TRACKER

☑ Project plan	☑ Collect complex data sets	☑ Analysis	☑ Folio of alternative designs	☑ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875



Used with permission of Tableau Software

FIGURE 5.5 Dynamic data visualisations can have a menu that allows the choice of data to be viewed. Interactive features include tooltip hover, menu buttons, year button and country button.

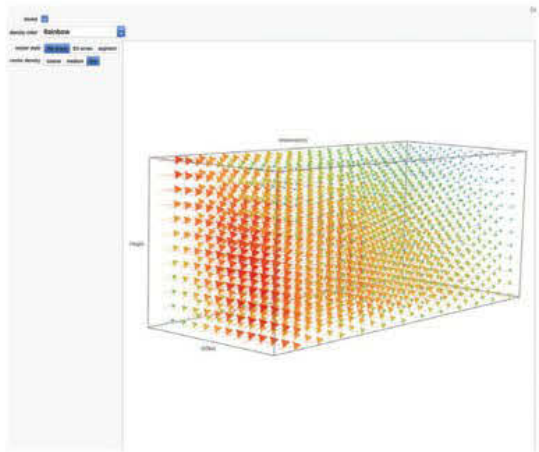
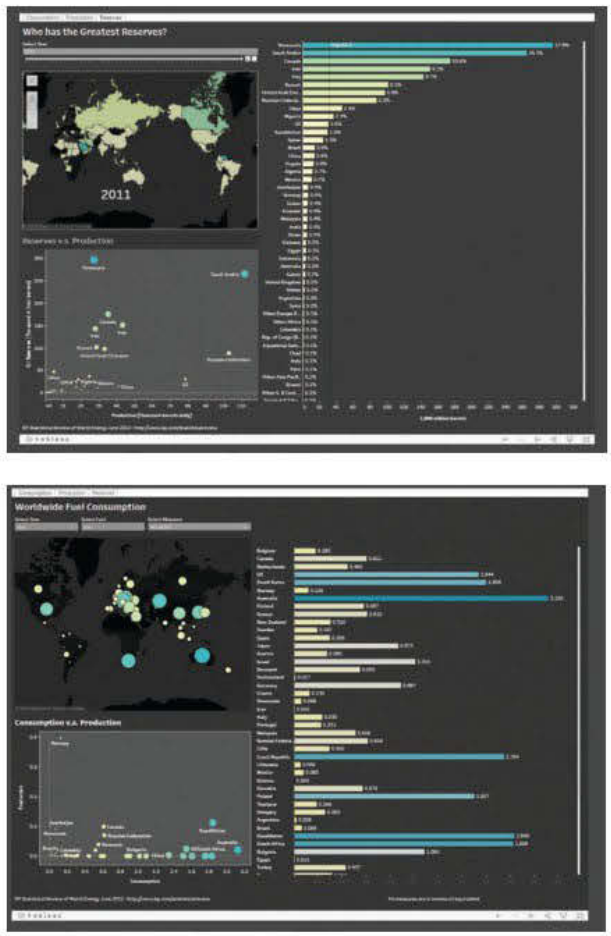


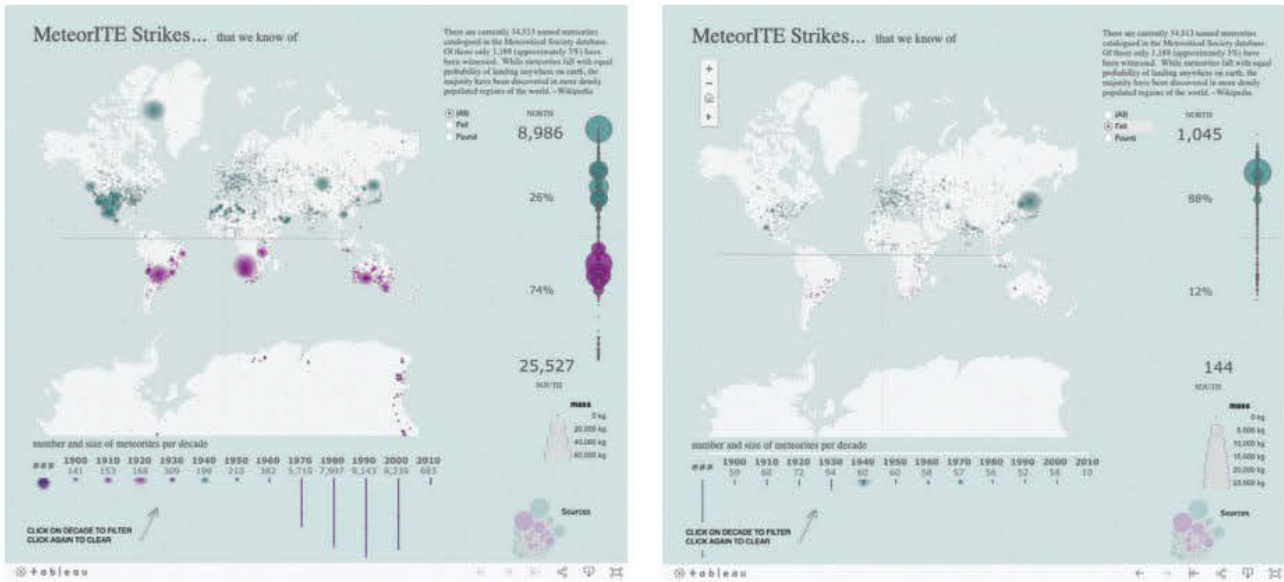
FIGURE 5.6 Interactive data visualisation for air speed near a circular fan (90°) in a model wind tunnel. The air speed near the fan blades is fastest, and slowest at the centre and edges. Menu choices alter the view displayed.

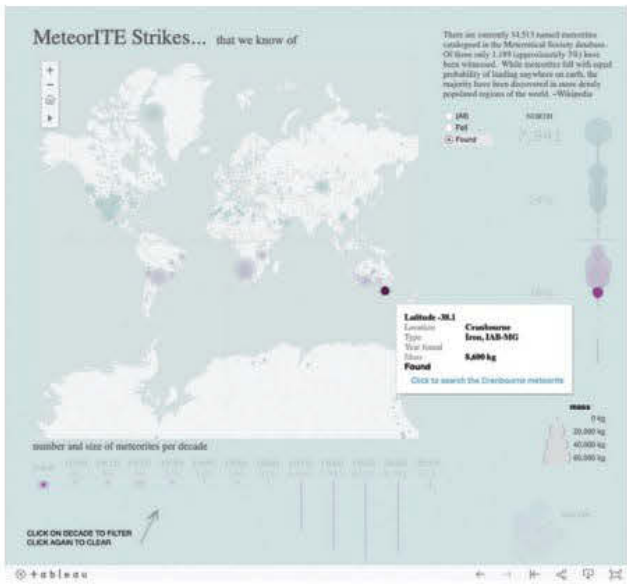


Used with permission of Tableau Software

FIGURE 5.7 Tableau Review of World Energy 2012. These two views illustrate how different displays can be chosen through a menu. Each element has hover tooltips to provide more detailed information about the country chosen. A country can be chosen by reference to the map, or to the country listing bar chart, or the scatter plot. Colour is used to indicate amount of resources.

9780170440875





Used with permission of Tableau Software

FIGURE 5.8 Meteorite strikes with radio button filter, tooltip hover, and URL links to data sources

Unit 4, Outcome 1

In Unit 3, Outcome 2, which was covered throughout Chapter 3 and Chapter 4, you devised a research question, and gathered data relevant to it to present findings that may support your question, or not. You provided complex data sets and information to justify your findings and supplied the specifications for creating your information. You referenced data sources, gave evidence of validation and preparation for the manipulation of data, and showed evidence of methods used to secure your information. You also provided your project plan in the form of a Gantt chart.

SCHOOL-ASSESSED TASK TRACKER

☑ Project plan	☑ Collect complex data sets	☑ Analysis	☑ Folio of alternative designs	☑ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

In Unit 4, Outcome 1, you will manipulate the data to derive information that will provide evidence for the research question. Then you will create an effective solution to present the findings from the research question. Unit 4, Outcome 1 makes up the second half of the SAT. In your graphic solution, you will communicate and substantiate a statement of your findings and whether your research question was supported, or to what degree the question has been answered.

Your solution should use an effective design and communicate the findings from your research question and your complex data sets clearly. It should be suitable for a nominated target audience and you should also use appropriate software tools.

During the development of your solution, you will validate input data, test it and evaluate its effectiveness at presenting your findings. You will apply your project plan to monitoring and recording your ongoing progress, and also assess the effectiveness of your project management strategy. The project plan can be modified to reflect changes.

This chapter will begin with a discussion about the characteristics of an effective infographic or dynamic data visualisation to help you understand how to deliver a solution that meets the requirements of the Outcome.

We will follow this by covering the formats and conventions, and then design principles, relevant to your solution. Finally, we will discuss software tools that you can use in the development of your graphic solution.

Procedures and techniques for managing files

Managing files safely and efficiently is a crucial skill when you are managing your project, especially when there are large quantities of valuable data on a computer, network or website. Chapter 2 introduced many of the factors to be considered when creating and managing files.

You will be managing many files as you develop your infographic or data visualisation solution and you will not want to lose valuable time and effort through ineffective file management practices.

Managing electronic documents forms part of an overall file management strategy. A **comprehensive management plan** will include all aspects of handling documents, including storage, retrieval, backups, archiving and security.

Computer operating systems and applications have wonderful search functions; however, these tools will not be as efficient as knowing where files are kept. An easily understood filing system allows you to go directly to the folder or file you need. Good file management practices always save time whether used by a single person on one computer or in organisations of any size with a few or hundreds of employees.

A management system

Establishing a management plan involves following these steps:

1. Creating the document management plan
2. Implementing the plan
3. Following through with the plan and establishing consistency

Creating documents

A business entity would have many types of documents. Sales estimates, email, spreadsheets, invoices and publicity brochures are created during the normal course of business.

What will be the process for creating these documents? For example: Are there standard templates with a logo letterhead for regular business documents, and where are they located? Is there a style guide to be followed? How are new documents date and time-stamped? How will documents be shared?

Naming documents

Clearly understood and meaningful filenames are always a good idea. Dates can also be included to ensure files are displayed in a sorted order. There are some limitations on filenames (seen in Table 5.1).

TABLE 5.1 Comparison of operating system filename limitations

Windows OS	macOS	Linux
Windows operating system (Win32) has a limit, called MAX_PATH, of 260 characters for the file path name. For example, C:\Program Files\filename.txt A filename and location may be acceptable until moved or copied to another location where the new path name exceeds 260 characters. Certain characters are reserved by the operating system and are not permitted for general use: • < (less than) • > (greater than) • : (colon) • " (double quote) • / (forward slash) • \ (backslash) • | (vertical bar or pipe) • ? (question mark) • * (asterisk)	There is no filename length limit, though after 1024 characters Finder has display issues. If the file is ever to be shared beyond the host computer, then 255 characters ensures compatibility with Win32 computers. There are only two characters that are disallowed by OS X: • \ (backslash) • : (colon) **Note:** Office does not allow : (colon) in filenames For cross-platform compatibility with Win32, limitations have become the default standard.	Just one character is reserved: • / (forward slash) There is no limit on filename length. For cross-platform compatibility with Win32, limitations have become the default standard.
These conditions may change – refer to the link for the current restrictions.		
Naming files, paths, and namespaces: Microsoft	Cross-platform compatibility: Apple	File-naming conventions in Linux: Linux

Some examples of naming documents can be seen here:

DataAnalyticsProjectV1.twbx has more meaning associated with the name than *DAV1.twbx*.

Subsequent versions can easily be identified:

- *DataAnalyticsProjectV2.twbx*
- *DataAnalyticsProjectV3.twbx*
- *DataAnalyticsProjectV4.twbx*

9780170440875

Version control is important when projects generate updates of a file.

In a similar way, including the date can indicate when the file was created, without opening the file or inspecting properties:

- *ElectricityNMIData20181201*
- *ElectricityNMIData20191201*
- *ElectricityNMIData20201201*

Storing and retrieving files within a directory structure

Managing files on a computer requires a consistent method of allocating locations of those files. Directories are the name given to the hierarchy of folders that can be constructed.

Every new installation of an operating system establishes a standard arrangement. Once a user begins creating folders and allocating files, the certainty of where a file may be found will be gone.

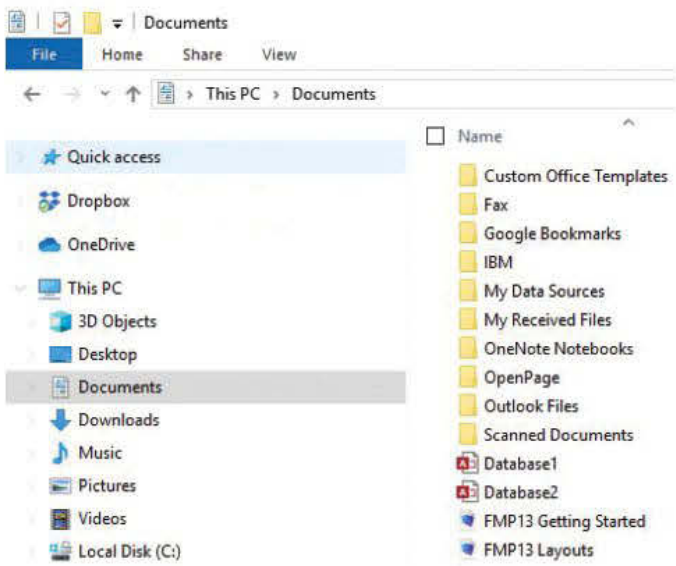


FIGURE 5.9 Windows 10 directory

Some basic file management approaches that apply to every computer file system are:

- All user created files should go into the MyDocuments folder (Windows 10) or Documents (macOS X).
- Brevity promotes clarity.
- Allow folder names to specify where files are to be saved.
- Keep folder names short.
- Separate work in progress from finished tasks.
- In your regular or monthly clean-up, consider moving files you no longer need to another drive, and consider calling that collection 'Archives'.
- Avoid deep and wide folder structures. If there are so many subfolders that the display is cut-off, consider alternatives.
- Keep similar file types together – applications, video, music, pictures, graphics, .docx, .xlsx, .pptx, and so on – since searching is made simpler.

9780170440875

- Limit the number of files kept; many files are unnecessary to keep once read and action taken.
- Your email Inbox is *not* a filing system. Choose to Delete, Move or Keep email as you read the message.

Where there are (too) many copies of a file, **version control** can be a problem. While one may be updated, the others are out of date yet retain the same name.

Use a shortcut and alias instead of creating multiple copies that may later become 'orphans' and never be updated.

To create a shortcut:

Windows	macOS
Right-click/Create shortcut (or Right-click/New/Shortcut)	Ctrl-click/Make alias

Once created, just drag-and-drop the shortcut to other locations.

Backups

There is a well-known 3-2-1 rule:

> *Keep three copies in two formats with one off-site.*

Clearly, if you have just one copy of a file, it is not that important!

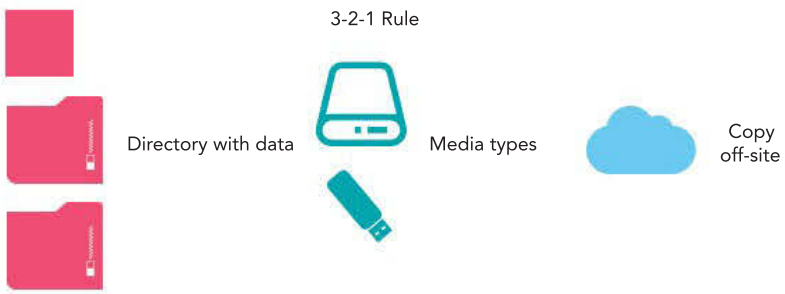


FIGURE 5.10 The 3-2-1 backup strategy

The purpose of a backup is to ensure a copy of the primary data is available when the original is damaged or lost. Data recovery will take time and the process and time taken forms part of a **disaster recovery plan (DRP)**.

The logic associated with this strategy reasons that keeping a copy of important data on a parallel storage device is not enough. Sometimes the D: drive may fail, so a different media is recommended. Another reason to have more than two copies of data is to avoid the situation where the primary copy and the only backup are stored in the same location. Cloud storage is now considered both a different medium and off-site.

Another time-worn statement is:

> *There are only two types of hard drive: those that have failed, and those that have yet to fail.*

9780170440875

RAID file storage is becoming more affordable. RAID (redundant array of independent disks) relies on the probability that 'only' one drive will fail at one time. Consider if all the drives are matched from the same manufactured batch, then a construction defect may cause the drives to be short lived. **MTBF (mean time between failures)** is a statistical expectation, and a dramatically shorter or longer time is possible. Reliability of drives is also measured using annualised failure rate (AFR) or as component design life (CDL). For consumer devices, a CDL would be five years and an AFR<0.8%. Alternative media includes SD cards, portable **hard disk drives (HDD)**, portable **solid state drives (SSD)**, CDs and **DVDs**. (Floppy drives are no longer considered a viable medium since access to a read/write drive may be unreliable.)

Bathtub curve

In all engineered devices, there is an effective serviceable lifetime where the performance is fit for purpose. Once the performance of the device degrades, the device must be replaced. Reliability engineering attempts to predict the working lifetime of a device by calculating the failure rate (symbol λ, called lambda, Greek letter 'el'). The MTBF is calculated as 1/λ

The 'bathtub curve' has three distinct sections.

1. Decreasing failure rate, or early failures usually due to manufacturing defects
2. Constant failure rate caused by random unpredictable events
3. Increasing failure rate due to the device wearing out

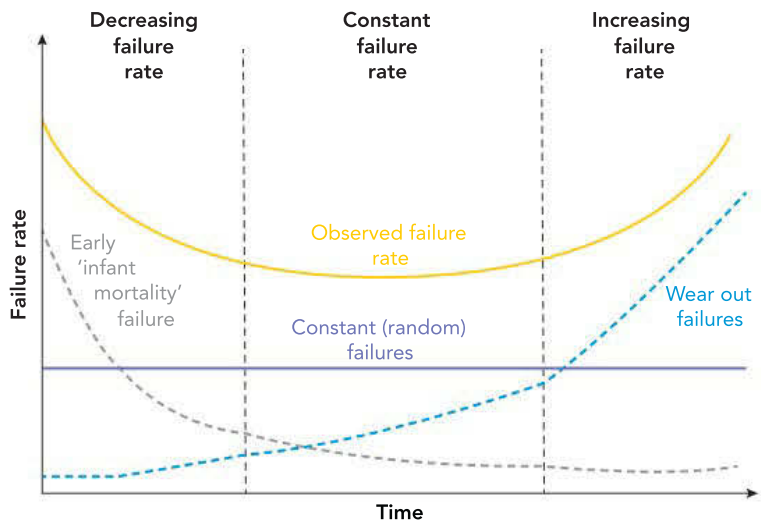


FIGURE 5.11 The 'bathtub curve' estimates the expected failure rate of an engineered device.

A tactic to reduce the first section of early failures is to 'burn-in' the device. This requires the device to be run under test conditions until the start of the second section. There are limits to this approach. This strategy is imprecise and has only been described after many thousands of drives have failed. The boundaries between stages are not well defined and is not particularly useful for predicting when a particular drive will fail, only that it has not failed yet.

Disposing of files

Placing a file in the recycle or trash marks the file to be deleted. The deletion takes place once the recycle/trash is emptied. The space on the drive is flagged as available. The next time a file is saved, this space may be overwritten.

THINK ABOUT DATA ANALYTICS 5.2

1. If a HDD is rated with an MTBF of 1.2 million hours, how long is this in days, weeks, months and years? If 1000 drives were operating for eight hours a day, how many would be expected to fail in the first year?
2. Research the drives in your computers. What is the expected MTBF?
3. What is the typical MTBF for a 256 GB micro-SDXC card X10?

9780170440875

Deleting a file does not remove the data from the drive or memory. When a file is 'deleted', usually the data remains untouched and the file allocation table (FAT) or table of contents (TOC) entry is removed. This will allow a 'file recovery' application to reassemble the data from the volume and the file can be read. The only change is that the name of the file may have been lost. Many file recovery applications regularly record the directory structure to enable a complete rebuild to be achieved.

To permanently delete a file, the data must be made unreadable – the data is 'scrubbed' or '**wiped**'. This will require the data to be overwritten with 0s and 1s. There are different levels of erasing data. A single overwrite of 0s and 1s is insufficient to remove the magnetic pattern of the original data. File recovery applications can recover data that has been overwritten. Multiple overwrites are necessary to prevent recovery. Three overwrites are considered to be sufficient scrubbing. This requires a cycle of writing all locations to read 1 then overwriting with 0, then repeated twice more, for three cycles.

Windows

SDelete is a Microsoft command-level utility that will securely clean the space on a logical disk. The number of passes can be specified.

macOS

The disk utility application that is integrated within macOS X has a function to erase 1, 7 or 35 times.

Disposing of a drive

When a computer is no longer to be used and/or allowed to be used by someone else, the existing data should be removed or deleted. The simplest strategy is to replace the drives with a new unused drive. If the drives are to be erased, then a secure erase process of overwriting at least three times must be completed. If the removed drives contain sensitive data, then the drive should be physically disabled or destroyed.

Archiving

A distinction to make is that backups are not archives. Archives are placed in medium- to long-term storage and are usually compressed to preserve storage space. Archives preserve a record of events and are often required by regulatory authorities. An index of files archived would assist locating the necessary document. For example, the Australian Tax Office requires all records to be kept for seven years, although prosecutions can go back 10 or more years. (Recall the purpose of a backup is to allow a recovery after damage or loss.)

Archived data may not be easily accessible to a regular user who logs in daily. While backups may occur hourly or daily, archiving usually occurs when files are no longer needed. For example, following the tax reporting season, after a major project is completed and signed off, when a semester unit is finished, or at the end of the year.

Factors affecting access of data

File access can be affected by the method of file organisation and the storage media onto which they are placed. There are different technologies available for storage, such as either mechanical or solid state drives.

File organisation and storage media

Frequently accessed files need to be placed in an easily accessible location on fast and reliable media. The choice of media for storage is often about capacity; however, other important attributes are:

- latency or speed of access
- reliability
- cost
- ease of use.

9780170440875

Use an internet search to locate typical latency times, access speeds and costs for:
- HDD
- SSD
- fusion drive.

Solid state drives offer the fastest access. The operating system, when placed on SSD, offers a typical start-up time of 10 seconds, compared with several minutes for HDD. Modern operating systems require at least 10% 'scratch space' for saving frequently accessed system files. A hybrid solution is a 'fusion drive' that has a small SSD attached to a HDD.

Functional capabilities of software

The features common to all data visualisation software will enable you to create a visualisation that communicates a clear message using images and text. Your final graphic solution will depend on your skills in choosing which elements to include. Just placing a graph into the software for display is simple enough. The challenge is to provide the 'best' graph or chart to convey your intended message. A 'dashboard' may be a useful method of presenting your findings.

Dashboards

Dashboards provide a summary of analysed raw data, usually on a single computer screen. This summary allows decisions to be made from the derived actionable knowledge. **Business intelligence (BI)** is a term applied to gaining actionable knowledge from information attained from analysed raw data.

Dashboards can be constructed with Excel, Tableau or Mathematica. Requirements for a dashboard are:

- 'clean' data where any duplications, errors or **null fields** have been adjusted
- a clear sense of purpose
- several support analyses that feed across into a summary chart.

The dashboard with multiple charts and text boxes often has dropdown menus, radio buttons, or other user interactions to choose the data to be displayed or the chart to be viewed.

The key to successful data visualisation is the storytelling. The purpose of the dashboard is not to overwhelm the audience with clever techniques; rather, it is to reveal the insight you have chosen that you want the audience to have. Your audience can explore the data with filters and interactive controls to further answer any questions they may have.

Your choices of software tools for the infographic or dynamic data visualisation will be guided by the VCAA requirements for both study and use for Unit 4, Area of Study 1. You are strongly advised to consult the existing requirements that are published annually in the VCAA Bulletin.



MyOnlineTrainingHub, www.youtube.com/watch?v=K74_FNnlIF8

FIGURE 5.12 Interactive Excel dashboard that summarises other worksheets into the one screen

An effective infographic or data visualisation

An effective infographic is one where the intended message is clearly communicated to the intended audience. No further explanations are required. Data visualisations often have more elements and frequently have interactive controls that allow the intended audience to change view settings. In these dynamic situations, an effective data visualisation would be self-explanatory. All the controls are able to be understood and manipulated. The controls may have unambiguous icons or are plainly labelled.



FIGURE 5.13 What constitutes an effective infographic or data visualisation?

Educating a target audience

In some respects, the way that you communicate your message can be as important as the message itself. As you develop your solution, you need to provide information that is suitable for your intended audience. This means that your content should:

- be gender inclusive
- be culturally inclusive
- apply common language
- be age appropriate.

9780170440875

The main thing is that you do not want to make anyone feel excluded from your audience or offended by your content, particularly on the basis of their race, ethnicity, nationality, language, culture, religion, attitudes, beliefs, customs, gender, sexual preference, gender identity or expression, appearance, dietary preferences, disability, illness, marital status, age, income, or education level.

That list may seem long and you may be concerned that it is going to be quite difficult not to exclude or offend anyone. Do not feel discouraged – just be considerate and sensitive. To a degree, you will already have an understanding of how to be sensitive when you communicate with others. All you need to do is broaden that understanding and apply it here. Consider the times when you have wished that others could be more tactful or sensitive with you or with someone else.

Gender

Your solution should be written with the assumption that your audience will be composed of females and males, as well as potentially intersex and transgender individuals. Excluding anyone from your solution on the basis of their sex (female, male, intersex), sexual orientation or gender identity (woman, man, transgender) would be unwise – and insulting.

Sexist language expresses bias in favour of one sex and treats another in a discriminatory manner. In most cases, sexist language benefits males at the expense of females. Using sexist language is considered insulting and patronising because it reinforces demeaning stereotypes and traditional gender roles.

It is not difficult for you to avoid using sexist language in your solution. Job titles that include a suffix such as 'man' or 'ess' to denote that each is a traditionally masculine or feminine role, respectively, are considered examples of sexist language, because they reinforce such stereotypes. Instead of 'police man', use 'police officer'. Instead of 'stewardess', use 'flight attendant'. Reinforce gender neutrality instead of adhering to stereotypes in your solution.

TABLE 5.2 Reasonable gender-neutral substitutions

Original	Gender-neutral substitution
Mankind	Humanity, humankind
Manpower	Workers, personnel
The man in the street	The average citizen, ordinary people
Sportsmanlike	Fair, sporting
Policeman, fireman, mailman, fisherman	Police officer, firefighter, postie/postal worker, angler
Actress, aviatrix	Actor, aviator
Man-hours	Worker-hours
Spokesman	Representative

9780170440875

RESEARCH

On 1 August 2013, it became unlawful to discriminate against a person on the basis of sexual orientation, gender identity or intersex status under federal law. Your solution, when being gender inclusive, should also keep this in mind. The Australian Human Rights Commission explains:

> Sex: a person's biological characteristics. A person's sex is usually described as being male or female. Some people may not be exclusively male or female ('intersex'). Some people identify as neither male nor female.
>
> Gender: the way in which a person identifies or expresses their masculine or feminine characteristics. Gender is generally understood as a social and cultural construction. A person's gender identity or gender expression is not always exclusively male or female and may or may not correspond to their sex.
>
> Gender identity: a person's deeply held internal and individual sense of gender.
>
> Intersex: people who have genetic, hormonal or physical characteristics that are not exclusively 'male' or 'female'. A person who is intersex may identify as male, female, intersex or as being of indeterminate sex.

- Refer to transgender people using only the pronoun of their identified gender, regardless of their biological sex.
- Never use the pronoun 'it' for people whose identified gender and/or biological sex is unknown. If you must use a pronoun, use 'they'.
- The most commonly accepted terms are now 'gay' and 'lesbian' for homosexual people.

Gender considerations

If you find yourself needing to use personal pronouns a great deal throughout your writing, and you need to maintain an even tone, you have two options. One is to use the third person plural pronoun, 'they'. This will allow you to protect the privacy of any respondents or subjects on whom you are reporting (including their gender), without dehumanising anyone with the universally unpopular 'it'.

Your other option is to use the correct pronoun on a case-by-case basis. If you need to discuss a particular interview subject's data somewhere quite specific in your solution, then you could refer to him or her as him or her, as appropriate.

Culture

Think of your solution as an environment or a space that your audience enters. In that environment, your audience needs to feel respected, welcomed and safe. In substantiating the results of your investigation, and justifying your findings, you should not harass or unfairly criticise or discriminate against anyone.

For example, using the vegan diet topic from Chapter 3 and Chapter 4:

> Do increasing numbers of people who live in South Melbourne eat a vegan diet because they believe it will give them health benefits?

It would be quite easy to turn the result of this research question into something exclusive and unfriendly, regardless of how well it was supported, if you had strong feelings about dietary preferences or if you disagreed with what your respondents and subjects thought. As a researcher, you are not being asked to share your personal feelings. It is important for you to remain objective. One of the best ways to help your audience feel safe and included is to maintain your objectivity so that you share only the truth of your findings.

Culture is made up of the shared ideas, customs and behaviours of a particular group of people in society, which distinguish it from another group in society. Cultures often create a sense of belonging and fellowship by excluding outsiders. The shared ideas, customs and behaviours of a culture can be quite expansive: from religious beliefs and shared sporting teams to dietary preferences and country of birth. Ultimately, your solution should not exclude anyone on the basis of their membership in a culture: whether they are Muslim, a Geelong Cats fan, vegan, or born in India.

Images

Your solution will most likely contain images (and perhaps video), and you need to be conscious of the possible effect these may have on your audience. If you are concerned that you need to use an image that may identify a child or a subject, or that could contain potentially offensive content, you could use pixelation (Figure 5.14).



Getty Images/Neil Mockford

FIGURE 5.14 Pixelation may be used to protect the identity of a child or a subject, or may be used to hide parts of an image that people may find offensive or distressing.

You must be sensitive when it comes to religion. Your target audience may be diverse on a religious level. Aboriginal or Torres Strait Islander People may find it distressing if a website includes images or names of deceased people. The portrayal of God in human form is forbidden by Jewish law. Representations of the prophet Mohammed can deeply disturb Muslims. Mocking or blasphemous images or cartoons of different cultures or religious leaders must be avoided.

You should avoid appropriating another culture's symbols, such as Native American headdresses, or reproducing another nation's flag, which could be seen as being particularly disrespectful. A nation's flag is an important symbol of that nation. Reproducing symbols that represent oppressive regimes, such as the Nazi swastika, will often cause anger and distress in many people.

Other factors that you need to be careful about include sexuality and violence. Some people are affronted by images of non-heterosexuality, sexual innuendo or full or partial nudity. Depictions of bloodshed, torture, injuries or violence could also cause distress.

Language

As the author of a solution for a wide audience, keep in mind that members of a different culture might not understand your vocabulary or references if they are embedded in your culture, but not considerate of your audience.

What constitutes swearing can be culturally specific. What seems harmless to you can be offensive to others. Humour is another ingredient best left out of a global presentation. Humour varies greatly from person to person, even within the same small cultural group, and what you find amusing may be considered inappropriate to others.

9780170440875

If you need to refer to another culture in your solution, be aware of potentially sensitive issues and always be cautious. If in doubt, report only the facts and omit your opinion. For example, if reporting information from the vegan hypothesis:

✗ 21% of respondents admitted trying a vegan diet because a TV advertisement conned them into it.

✓ 21% of respondents reported trying a vegan diet in response to TV advertisements or other promotions.

In the first example above, the language shows clearly through words such as 'admitted', 'TV advertisement' and 'conned' that the writer has an attitude towards the respondents – that they are gullible. More than that, the word 'admitted' has the connotation of being a confession, which suggests that the writer feels the respondents have done something *wrong*. Any members of the vegan culture – people who eat a vegan diet themselves – would feel excluded by that language. However, the second example uses more neutral language, 'reported', 'in response' and 'TV advertisements or other', which omits the writer's personal feelings.

If you need to discuss other cultures in your solution, be aware of potentially sensitive issues and always be considerate. Report the facts – not your opinion.

Rather than using euphemisms in your writing to avoid shocking or offending people, use correct formal language. As a researcher, you need to present the facts, so it is not appropriate to use euphemisms, such as 'passed away' to mean 'died', 'expecting' to mean 'pregnant', or 'facilities' to mean 'toilet'.

Commonality of language

English is often considered a global language, but some of your global audience may not be as fluent in English as you are. Some speakers of English in your audience will not be native speakers, and some speakers may speak American English, British English or other dialects of English. Your choice of vocabulary should take this into consideration and be easy to understand.

- Replace an obscure word with a simpler word or phrase that means the same thing, but try not to sound **condescending**.
- Keep any text sentence short, with one main idea per sentence.
- Punctuate clearly. Obey conventions.
- Have someone proofread your work to find spelling errors, culturally bound ideas or vocabulary that may be confusing or incomprehensible, offensive expression, lack of clarity, repetition or ambiguity.
- Be careful using colloquial language – the informal language used in conversation with peers within a culture. Although it seems natural to you, you may not recognise that it could be unintelligible or misleading to other cultures.

Misunderstandings between speakers of different languages may end in confusion, or worse. Even native English speakers understand some words to have different meanings.

For example, to an American, a 'bum-bag' would be a sleeping bag for a homeless person, but to an Australian, it is a piece of apparel. To a New Zealander, a 'dairy' is a milk bar (Figure 5.15), but to an Australian, it is a place where cows are milked or where milk is processed.

9780170440875



Alamy/Greg Balfour Evans

FIGURE 5.15 A 'dairy' in New Zealand is what we call a 'milk bar' in Australia.

A car that runs on gas in Australia uses LPG, but in the USA, 'gas' is petrol. Scallops and flake (Flake) in New South Wales mean potato cakes and a chocolate bar. In Victoria, scallops mean shellfish and flake means shark fillets dipped in batter and deep-fried. Australians know thongs as casual summer footwear; Americans only know thongs as an item of underwear. If you have been to the USA, you will know that you enter a building at the first floor while in Australia you enter a building at the ground floor. To top it off, being caught outside in your pants is quite funny in British comedy shows because to the British 'pants' are underwear. Australians do not see what is so funny since we consider pants to be trousers or jeans.

In your visual solution, you may not realise that you are using terms that your readers do not know. Obtaining feedback from someone else may highlight words, expressions or references that should be changed or explained.

Age appropriateness

Try to make your writing as accessible as possible to readers of different ages. Your target audience may include children as well as adults. You are not expected to be able to cater to everyone: it would be impossible because of the wide and unpredictable range of your audience's vocabularies, life experiences, comprehension skills and areas of interest. However, you can make an effort to write for a general audience and be aware of techniques you can use to cater to some age-related needs.

You can calculate an estimate of the age or education level a reader would need to understand your writing. Microsoft Word can give you readability statistics of your document. Click the round Office button (or File, in older versions) > Options > Proofing > Show readability statistics (which should be a checkbox) (Figure 5.16).

Readability

Flesch Reading Ease	61.3
Flesch-Kincaid Grade Level	7.1
Passive Sentences	0%

FIGURE 5.16 Readability statistics for the sentence 'Vegan dieters eat grains, legumes, nuts, seeds, vegetables and fruit.'

When you perform a spellcheck, Microsoft Word shows a rating of how readable the document is. Two of the readability measures, 'Flesch Reading Ease' and 'Flesch-Kincaid Grade Level', are calculated according to the number of syllables and words in a sentence; essentially, the shorter your sentences are, and the fewer the syllables, the easier they are to read, especially for younger audiences. Generally, aim for a higher Flesch Reading Ease number and a lower Flesch-Kincaid number.

The passive sentences percentage is included as a measure of readability because the passive voice is less effective and clear in writing than active voice. The higher the passive sentence percentage, the less readable your document is to most of your audience – young

9780170440875

and old, native English speaker or not. The active voice is more effective in writing because it is simple and direct.

✓ ACTIVE: Vegan dieters eat grains, legumes, nuts, seeds, vegetables and fruit.

✗ PASSIVE: Grains, legumes, nuts, seeds, vegetables and fruit are eaten by vegan dieters.

TABLE 5.3 Breakdown of active and passive voice in simple sentences

	Subject	Verb	Object
Active	Vegan dieters	eat	grains, legumes, nuts, seeds, vegetables and fruit.
	Agent (the doer)	the action (doing)	affected (done to)
Passive	Grains, legumes, nuts, seeds, vegetables and fruit	are eaten	by vegan dieters.
	Affected (done to)	the action (doing)	agent (the doer)
Active	The vegan diet	excludes	meat, dairy, eggs and seafood.
	Agent (the doer)	the action (doing)	affected (done to)
Passive	Meat, dairy, eggs and seafood are excluded		by the vegan diet.
	Affected (done to)	the action (doing)	agent (the doer)
Active	The majority of respondents (56%)	preferred	the non-dairy aspect of the vegan diet.
	Agent (the doer)	the action (doing)	affected (done to)
Passive	The non-dairy aspect of the vegan diet	was preferred	by the majority of respondents (56%).
	Affected (done to)	the action (doing)	agent (the doer)

Writing for younger readers

When writing for younger audiences, care must be taken to provide information at the appropriate level of reading ability and suitable maturity. Consider using simple graphics to convey your message.

- Create an engaging discussion.
- Use a simple sentence structure.
- Make clear points. Younger children are literal-minded and do not yet reason abstractly or understand irony, sarcasm, analogies, ethical dilemmas, implied meaning or metaphor. They may not detect subtleties.
- Choose from a suitable vocabulary, but do not be too concerned with introducing new, useful words.
- Explain any technical terms in simple language.
- Be careful of discussing adult themes, such as sex, violence, drugs, alcohol or death.
- Consider the type of illustrations for younger readers. Cartoonish graphics may be appropriate. Slightly older readers may appreciate simple line illustrations, or graphics – or none at all. Clip art has become a little clichéd, so this type of illustration is not recommended.
- Avoid being patronising. Younger readers are inexperienced, not unintelligent.

9780170440875

Writing for senior readers

Writing for elderly/older/senior readers also has its challenges, but it is unfair to generalise about senior people's technological expertise. For every older person who asks for larger text on-screen or simpler instructions, there is another older person who can move swiftly using regular equipment and eagerly embraces new technologies.

In general, on a technical level, be sure that any text is easy to read for those whose eyesight may not be as sharp as younger readers, and is relatively easy to navigate physically. Authors writing for older readers should remember the following.

- Try not to use too many similar colours, nor to obscure text by background colour or images. As eyesight declines, shades of blue can appear faded or desaturated.
- Avoid large blocks of white text on dark background. Ideally a short paragraph of darker font on a light background is more effective.
- Always provide 'alt text' descriptions for images, consider providing a full text-only version of your infographic for vision-impaired audiences, or provide a HTML/CSS version so that a screen-reader can access the content.
- While audio is not usually critical to interacting with a solution, be sure to remove background noise if you are recording your own evidence.
- You could test your solution using screen-readers, which convert text into speech for people who are vision-impaired. For macOS: use the built-in VoiceOver app. For Windows: Download the NVDA app. NVDA supports Microsoft Edge, Firefox and Chrome browsers.
- Provide subtitles or captions if video is essential to your information.
- Do not put controls (such as buttons) too close together. Make controls a generous size. Include text or captions that describe their function.

THINK ABOUT DATA ANALYTICS 5.3

Research 'infographics accessibility' and summarise what can be done to make an infographic accessible.

Accessibility

When developing your graphic solution, consider how to communicate your message more effectively to people with special needs. Avoid making communication unnecessarily difficult for people with special needs, including the young, the elderly and those with disabilities, such as low vision, colour blindness, limited mobility or limited language or computer skills.

Some of the measures you can take have already been listed in the 'Writing for younger readers' and 'Writing for senior readers' sections previous. Other suggestions are described in Figure 5.17.

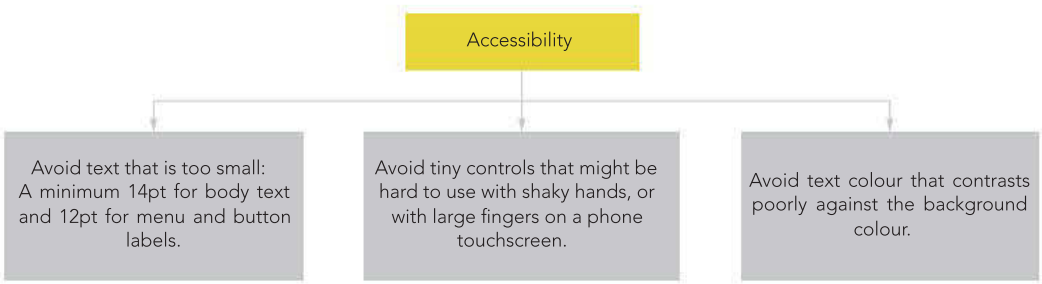
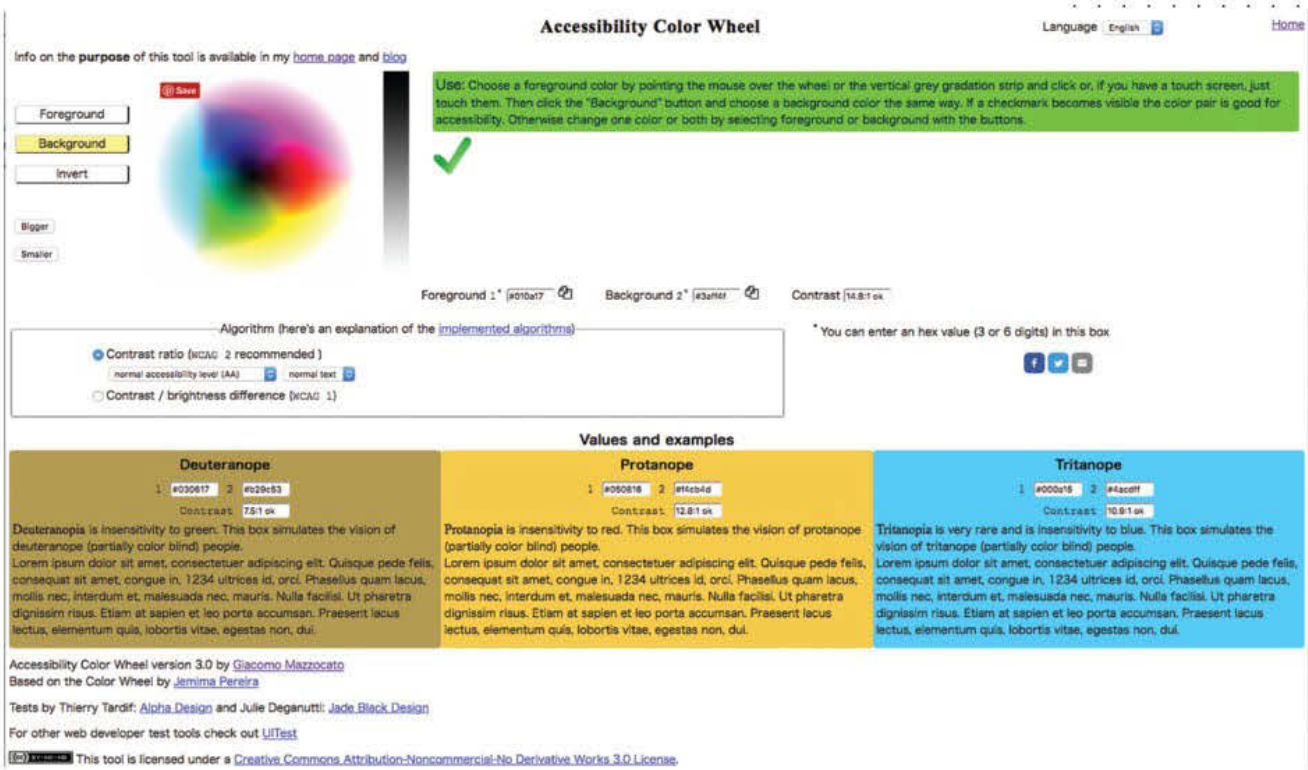


FIGURE 5.17 Suggestions for accessibility (also refer to design principles)

Avoid using text that is too small. As a guideline for infographics, use a minimum of 14pt for body text and 12pt for menu and button labels – larger, if possible. Avoid using very small controls that might be hard to click easily with shaky hands, or with large fingers on the small touchscreen of a smartphone. Avoid text that contrasts poorly against the background colour. For example, avoid white text on dark backgrounds.

Refer also to the design principles for accessibility, error tolerance, ease of use and navigation later in this chapter and Chapter 4.

Interactive colour accessibility checker will verify how your choice of colours will be viewed



Content created by Giacomo Mazzocato. Used with permission.

FIGURE 5.18 Accessibility Color Wheel is an interactive web page that tests the chosen colour for issues created by foreground and background contrast effects.

Clarity

As already discussed in Chapter 2, whether spoken or written, your solution's language should be clear enough for most audience members to understand. If you cannot convey why you concluded that your research question was supported and justify this using multiple types of data, your solution will fail.

The advice previously given regarding explaining technical terms and using direct language especially applies here. For infographics, it is a good idea to use short sentences, short paragraphs and dot points. A graphic solution written entirely in dot points and one or two sentence paragraphs can be very engaging if it draws the audience *in* using a combination of colour, images/graphics and text positioning.

Could a well-chosen image replace some of the text? You will need to use careful paragraph placement here and there to develop an argument, sustain a narrative and engage your audience's interest. Use straightforward phrasing. If you are concerned about needing to use longer paragraphs and are worried about long blocks of text, break them up with

5.4 THINK ABOUT DATA ANALYTICS

How could you test your solution for clarity?

SCHOOL-ASSESSED TASK TRACKER

☑ Project plan	☑ Collect complex data sets	☑ Analysis	☑ Folio of alternative designs	☑ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

subheadings, tables or relevant illustrations as needed. If you need to use a long sentence that worries you, look at the punctuation: can it be broken into multiple sentences? If not, can it be framed with shorter sentences on each side so that it is easier to read? If you are still uneasy, revisit the sentence and ask yourself: is it the right sentence?

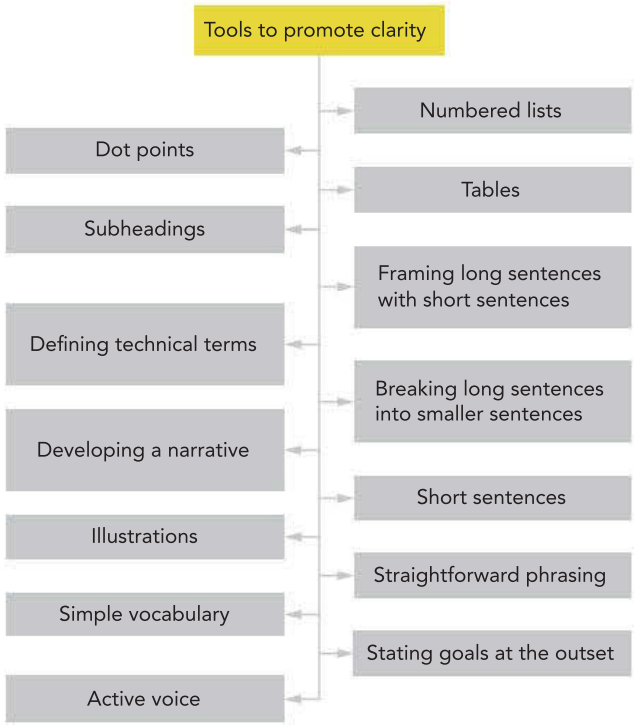


FIGURE 5.19 Tools to promote clarity

Readability

Readability refers to the ease with which a reader can understand the content of a written text. Readability can be affected by typeface and the use of white space. Both of these elements were covered in Chapter 2, page 91.

Relevance

Relevance refers to staying on topic. Make sure that you only include material in your solution that is directly related to the topic that you are researching. For example, if you are researching the popularity of the vegan diet among residents of South Melbourne, you do not want to start presenting information about a vegetarian diet among those residents.

Accuracy

Your infographic or data visualisation solution aims to educate your audience. Make sure that the information you provide is correct based on the evidence you have gathered from investigating your research topic. Any data sources you used during Unit 3, Outcome 2 should have been reputable sources. See Chapter 2, page 105 for a further discussion of accuracy.

Usability

As a writer of your research solution, you face many difficulties in reaching larger audiences. There is such variety in the hardware and software used by audiences that it may be difficult to create a single solution that is equally usable by everyone. There are several potential factors to consider, including plug-ins, browsers and hardware.

9780170440875

Plug-ins

Plug-ins are software modules added to applications to enhance their functionality, such as browser plug-ins to block pop-ups or to display PDF documents.

If you use a Flash animation to convey vital information, but your audience using iOS mobile devices cannot access Flash media, your message cannot be communicated. To prevent this problem, there are other options:

- Produce one version of your solution on a website using the plug-in, and another version without it. This doubles your development and maintenance effort.
- Produce a very basic solution website that every browser can render properly.
- Ignore audience members who lack the plug-in.
- Produce a solution website that uses a different plug-in that is more likely to be supported by a wide range of platforms, such as JavaScript. This still may ignore some audience members who lack the plug-in.

The decision is not easy to make, and such problems are ongoing because of changing technology. The most you can do is make smart choices based on the current technology of your likely audience.

Another nail in Adobe Flash's coffin – Google Chrome to block Flash

Amazon puts another nail in Flash's coffin

Final nail in the coffin for Flash?

THINK ABOUT DATA ANALYTICS 5.5

Adobe Flash was once the dominant media technology for website animation. Research into why it has lost support from major technology companies, such as Apple and Google.

Browsers

When constructing your infographic or dynamic data visualisation, you may choose to display the final presentation through a browser. Excel, PowerPoint, Tableau and Mathematica (through Wolfram Cloud) all have the capability to publish through the web. There is a clear advantage in this approach since the intended audience is not required to install the application software on their device. The World Wide Web Consortium (W3C) publishes standards that websites and browsers are supposed to obey to be compatible with each other. However, no browser has perfectly achieved this aim and some sites do not work well in some browsers.

World Wide Web Consortium (W3C)

If your graphic solution uses a browser, you may have a problem. As the developer of the graphic solution, you cannot know how your audience has configured browsers and devices. Are pop-up windows disabled? Are ads blocked? Is sound muted? Is JavaScript blocked? Is image loading turned off? Are cookies disabled? Any of these can cause your graphic solution to perform less adequately than expected.

Try to ensure your solution is usable for as many audience members as possible.

- Test your research solution in all the dominant browsers, including the current versions and the most recent old versions.
- Use online services that perform compatibility checks of your solution.
- Try to anticipate common problems, such as disabled Flash or cookies, and provide back-up methods to convey information. If audience members set images not to auto-load, use alt text to describe missing images. Provide email contact links in case feedback forms do not work.

9780170440875

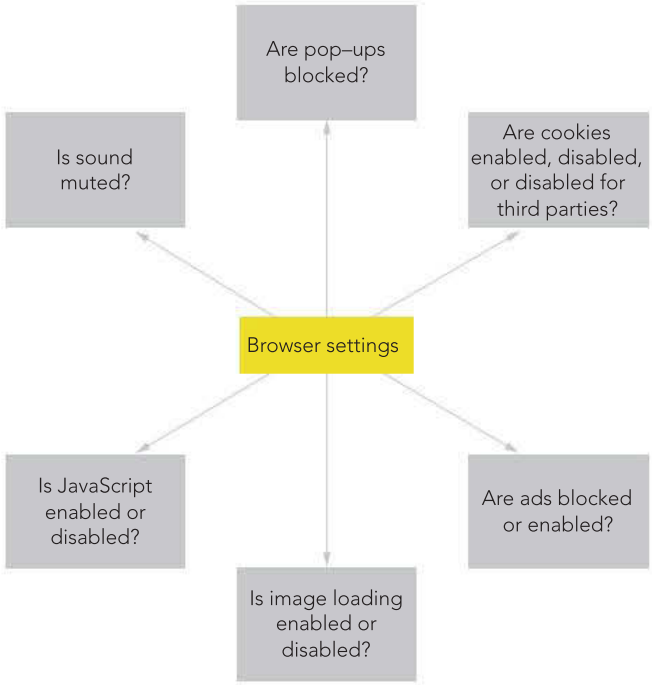


FIGURE 5.20 Browser settings will differ throughout your audience. Before you finish your solution, you will not know what browser settings your audience will have chosen, but these are factors to keep in mind.

Hardware

If you build a graphic solution using a 1920 × 1080-pixel monitor, you may be disappointed to see how the page renders on the 5-inch screen of a smartphone (Figure 5.21). Catering equally for all devices is difficult. Not everyone has the time, skill and budget needed to create and maintain both mobile and desktop versions of their solution, or write the JavaScript that can cope with browser differences. However, you could try to cater for the most popular current devices – smartphone, tablet and desktop – so you can communicate your message to as large an audience as possible.



iStock.com/milindri

FIGURE 5.21 The same site viewed on four very different displays: a smartphone, a tablet, a HD monitor and a laptop

9780170440875

Media

Avoid media that is only playable on certain platforms, such as macOS, or with certain players, such as Windows Media Player. Media not supported by major technology suppliers, such as Flash, should also be avoided. Restrict yourself to standard media types (Table 5.4).

TABLE 5.4 Standard recommended media formats

Media usage	Recommended format
Images	JPG for photos GIF or PNG line art and logos
Audio	MP3 WAV gives perfect quality because it is uncompressed; however, it is 10 times larger than MP3
Video	MP4 (DivX or FLV are not universally supported)

Not all computers, smartphones and tablets can read all media formats. Some players (such as VideoLAN's VLC media player) can play almost anything; others will play only some media.

Timeliness

Timeliness is important to ensure that users do not experience significant delays when viewing your solution due to delays in loading media-rich and multimodal files. See Chapter 2, page 105 for a further discussion of timeliness.

Another meaning of timeliness for the purposes of your graphic solution refers to the actual currency of your information. Refer to the 'Data integrity' section in Chapter 3 (page 139) for a further discussion.

Completeness

As already discussed in Chapter 2, completeness means that the information you are presenting in your infographic or data visualisation is just that: complete. Ensure that you present all of your findings to support the conclusion you reached in Unit 3, Outcome 2. If you do not do this, you will not be able to convince your audience successfully that your research question was genuinely supported. Although you have a degree of latitude in how you present the information you have obtained from gathering and processing your data, you still must present all of it so that it can be considered properly. You can use some of the tactics referred to in the 'Data integrity' sections in chapters 1 and 3 as a guide if you need further assistance.

Attractiveness

You need to ensure that your website has visual appeal or attractiveness (see Chapter 2). The more attractive users find your solution, the longer they will view it and trust its message.

Examples of attractive websites

Keep it simple:

- Give your solution a consistent look and behaviour.
- Make interfaces clean, simple, easy to learn, easy to use and attractive to look at. Limit the number of colours and typefaces used.
- Avoid complexity. No-one will be impressed that you just discovered animated GIFs or used 200 typefaces.

9780170440875

- Use subtle colours. Neon or overly bright colours are difficult to tolerate for any length of time.
- Employ leading, kerning and white space (see pages 91–92) to make text more readable.
- You should generally obey conventions. Doing things your own way just to be clever or different, without a substantive reason, annoys an audience.
- A poor interface is confusing, hard to interpret, and causes user errors. It discourages your audience from staying long enough to receive the message you aim to communicate.
- Consider accessibility needs. For example, attach alt text to images so vision-impaired users with screen-readers can hear what the images represent.
- Anticipate common user errors and cope with them tolerantly, such as always providing a 'Back' or 'Cancel' facility so users are not locked into a path from which they cannot escape. This relates to design principles > functionality > accessibility > error tolerance.
- Be informative. Provide indications of how well a long operation is proceeding, such as how much has been downloaded or calculated so far, so that users do not think the computer is frozen or a website connection has crashed.
- Let typical end-users beta test your interface and pay attention to their feedback.
- Use the most appropriate graphical user interface (GUI) controls. For example, to force a choice of a single item from a selection, use radio buttons. To allow zero or multiple choices, use checkboxes.
- Be flexible. Your audience should be able to choose from menus, hyperlinks, dropdowns or buttons to perform an action. Everyone has their own preference.

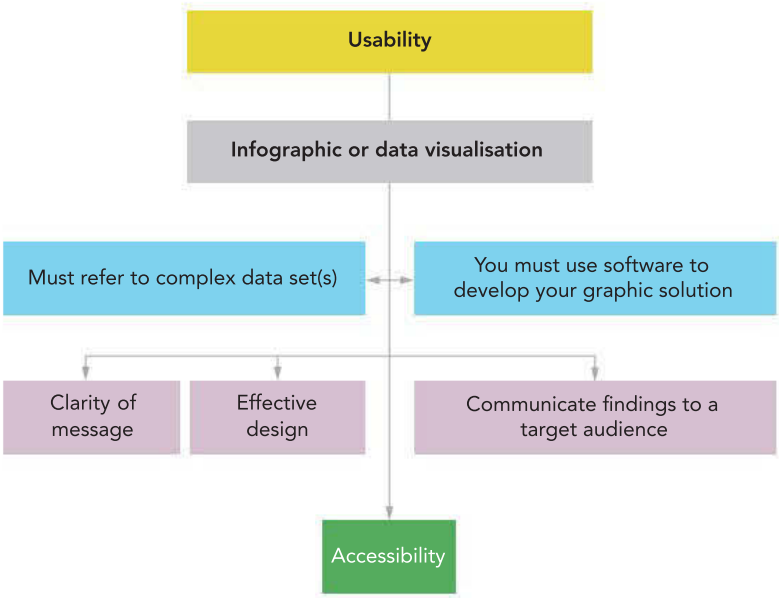


FIGURE 5.22 Requirements for the data graphic solution

Manipulating data

Explore and learn the features that the processing software you have chosen offers. It will take you a long time, and you will likely produce a subpar solution if you do not take time to become familiar with your software before you start to build. Examples of readily available software are Microsoft Excel, Mathematica and Tableau. While Python and R are also available, the amount of prerequisite skills and knowledge required to generate any useful end-product may not be possible in the time available for this section of the course.

Remember, as discussed at the beginning of the chapter, you need to choose software that allows you to perform a number of tasks, and you need to perform those tasks in your solution. Figure 5.22 restates the requirements for your authoring software.

9780170440875

You can use several types of software tools to satisfy these criteria, such as the following.

- Microsoft Access and Excel (2019 and Office 365) have add-ins that provide additional data manipulation features – notably geographic presentations when latitude and longitude data is provided. Output can be sent to PowerPoint, Word, Tableau or displayed in HTML format to a web browser.
- Slideshow software applications, such as Microsoft PowerPoint, Keynote (for Mac) or Sliderocket. There are services to upload and share PowerPoint presentations, such as Powershow.com and Microsoft's PowerPoint add-on Office Mix.
- Online slideshow presentation makers, such as Google Slides, Microsoft Sway and Prezi, take images of all graphs and arrange as a photo gallery to display the filtered data on demand. **Note:** Data cannot be updated, but display is interactive.
- Tableau desktop is available under an education (student) licence. Your teacher can acquire a registration code for the class. Tableau provides comprehensive data manipulation, selection and display capabilities. Multiple data sources can be accessed as either snapshot (static) or updated (dynamic). **Note:** The free version posts your saved file to a public gallery. Tableau Viewer allows others to view the files.
- Flourish is an online data visualisation tool, and the trial is free.
- Zoho Office Suite has database and spreadsheet capabilities.
- Google Data Studio has data visualisation capabilities, though it has a limited range of charts.

Google Sites
Zoho
Tableau

Heading styles

Cell styles, such as those used in Microsoft Excel allow authors to describe formatting styles once, use the style many times, and change styles easily.



FIGURE 5.23 A data visualisation website created using Google Sites

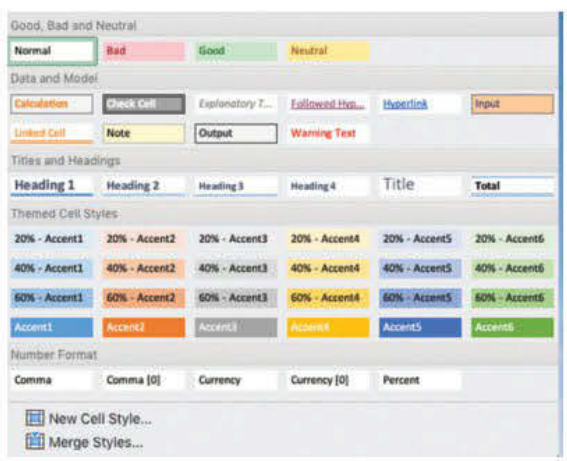


FIGURE 5.24 Heading styles in Microsoft Excel showing some in-built styles (such as Heading 1) and the option for adding some custom-made styles

In a large worksheet with several pages, headings and subheadings, it is easy to change your mind about formatting as the document grows. Going back to each cell and reformatting its text size, colour, alignment, typeface and style is exhausting. However, with cell styles or CSS, you can redefine any heading or text style throughout the worksheet instantly.

SCHOOL-ASSESSED TASK TRACKER

☑ Project plan	☑ Collect complex data sets	☑ Analysis	☑ Folio of alternative designs	☑ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

Customise your tools

Straight out of the box, multimedia authoring software is set up to suit the average user. However, an advanced user will customise tools to maximise efficiency. Most sophisticated software will allow you to add and change keyboard shortcuts, menu ribbons and default values, such as typefaces and line spacing, to suit your preferences.

Learn your tools

Spend some time browsing the submenus of your favourite software. You would be pleasantly surprised by the power and assistance that is available. Microsoft once surveyed users of Word and Excel, asking what features they would like to add: 90% of the requested features were already in the program, but users had not found them.

Templates

It is common to create many documents that follow the same pattern using, for example, the same logo, links, margins, colour schemes, header/footer and copyright statement. Instead of recreating all those settings for every new document, create a blank template document with those settings already entered and save it. Whenever you need a document of a similar type, load the template document, add the new content and save with a new name. This saves time, ensures you do not forget to add necessary components, and promotes a consistent look and feel for all of your related documents.

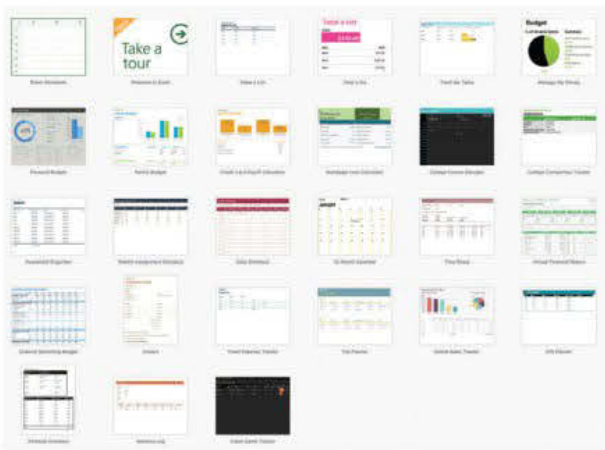


FIGURE 5.25 New spreadsheet templates in Microsoft Excel

A slideshow template may consist of a few special slides (a title slide, a slide for contents, one for your references and a concluding slide) plus one slide that will be the parent of all of the slides containing body text.

Pre-made templates of worksheets, slideshows or other infographic documents may be available to you with pre-designed design elements. In your graphic solution, you should not use templates created by other people, since this will not be demonstrating your ability to choose and apply design principles.

9780170440875

Formats and conventions

Presenting your information effectively requires a knowledge of both formats and conventions. You have already encountered this topic in Chapter 2.

A **format** is the form in which information is presented, such as a web page, pie chart, text in paragraph, table, comic strip, limerick, pop song or newspaper article. There may often be several formats from which to choose, such as presenting data in either a table or a graph.

Conventions are the accepted techniques that an audience will expect to find when a format is used. Each format has its own conventions.

TABLE 5.5 Formats and conventions relevant to your graphic solution

Format	Conventions for format
Pie chart	• Coloured or patterned slices • Largest slice starts at the 12 o'clock position, and slices proceed clockwise in order of decreasing size • Slices are labelled or the chart has a legend • Limit the number of slices to around six, unless data values are included
Cartoon	• Speech bubbles have a pointed tail to the speaker's mouth • Thought bubbles have a trail of bubbles leading to the thinker's mind; other characters cannot hear these thoughts • Time moves in frames from left to right, top to bottom
Web page	• Underlined or coloured links for URLs • Site logo acts as a link to the homepage • Landing page is named index.html, index.htm or index.php • One main topic per page • Large pictures are linked to small thumbnail versions • Image formats are JPG, GIF or PNG • A contacts page lets visitors contact the site owner • Sites have a search facility • Up to 4–5 clicks should take users to 80% of the documents they want to view • Sound and video should never play automatically when a page loads
Table	• Column headings are bold and in the top row • Any units appear in the column heading • Numbers are right-justified or centred on the decimal point • Australian currency has two decimal places • Border lines separate columns • Totals appear at the bottom row and/or rightmost column
Text	• Only centre short units of text, usually left justified • Sans serif font for on-screen body text to ensure readability • Serif heading text for visual distinction • Consistent fonts, type sizes (usually 9–12 pts) • Lines are no longer than 60 characters (10–11 words) • Limit paragraph size (4–8 lines) • Use wide margins • Avoid use of all capitals • No underlining – use italics or bold for emphasis • Consistent heading size and styles • One space at the end of sentences

9780170440875

You should follow established conventions because information is far quicker and easier to find and understand if it is presented in a predictable manner. You should not force your audience to learn new ways of information presentation for no reason. Audiences generally expect conventions to be used, though consider that conventions can also come in different strengths.

If you disregard **mandatory conventions**, you are breaking the law. For example, whether a country chooses to drive on the left side of the road or the right is quite arbitrary. However, once a nation has chosen a side, people who choose to drive in the incorrect lane will be punished for disregarding legal conventions.

Preferred conventions are those where most people have a distinct preference about how something should be done. For instance, in a large document, people expect to find page numbers. In reference books, readers would be upset if they found there was no index. Hyperlinks in a web page are traditionally underlined.

Optional conventions offer you a real choice. Some individuals may prefer to see a book's page numbers centred at the bottom of the page, but would not write angry letters to the publisher if page numbers appeared at the top of the page. Similarly, some individuals prefer to see a link to the site map at the header of a page rather than at the footer.

If you do not learn, or do not follow conventions, you will hinder the communication of your message and possibly tarnish your credibility with your audience. To maximise the readability of your final presentation, and to better communicate its message, discover the conventions that apply to the format you choose to present your solution and use those conventions consistently and accurately.

Manipulating data with software

To a computer, everything is data: text, music, programs, movies, photos, games and web pages. They all need processing to produce a solution. In Unit 3, Outcome 2, you needed to be competent in data processing to manipulate the data to generate information to provide evidence for your research question. In Unit 4, Outcome 1, you need to manipulate data such as images, video, charts, text and audio effectively to develop your infographic or dynamic data visualisation.

Techniques for creating your graphic solution

After data has been manipulated and your findings are made, you should use the best elements from those charts, images and summary statistics that support your findings. These will now be assembled to tell a narrative or story.

Previously in Chapter 4, you analysed your data set and revealed insights and trends that formed the basis of your research findings. This section will develop the clarity of those findings and provide visual support for your research statement.

You may have already used data manipulation software, Microsoft Excel, Wolfram Mathematica or Tableau.

The software you choose to create your graphic solution, which must be either an infographic or a dynamic data visualisation, will use a combination of data manipulation software and presentation software. (**Warning:** While all three data manipulation software also offer a presentation mode, learning how to assemble within Mathematica and Tableau may require time well beyond that available for this course.)

9780170440875

Consider instead that graphic elements can be created in data manipulation software and transferred into the presentation software. You can arrange to have dynamic changes by introducing simple 'switches' to choose a view of the data that corresponds with a user choice. The displayed static image changes in response to a menu choice.

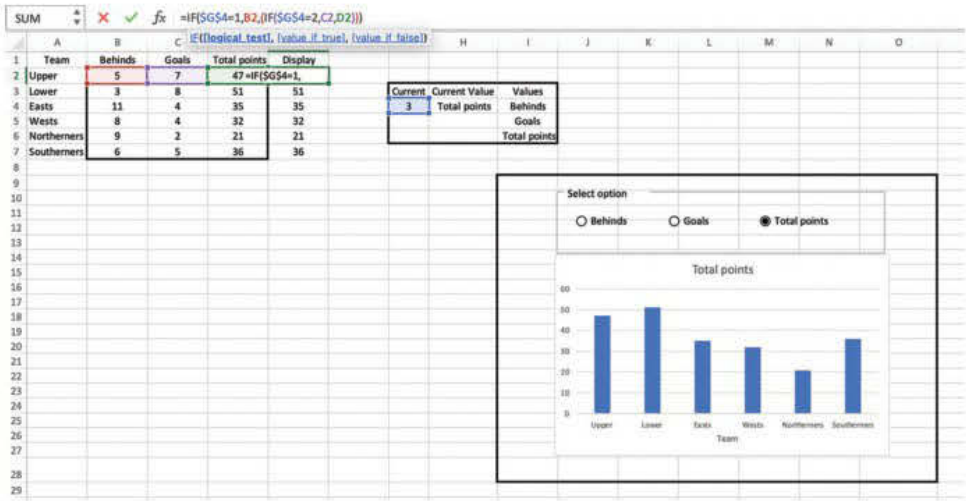


FIGURE 5.26 An example of a dynamic Excel chart with radio button to provide user interaction to choose the chart viewed

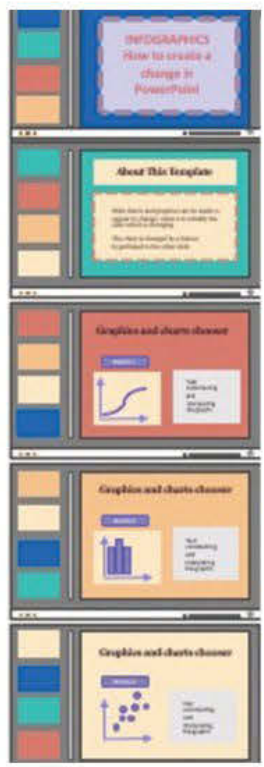
PowerPoint and Keynote are well equipped to achieve this task simply using hyperlinks within the presentation. Each of the different graph options are recorded as an image and a menu provides a range of choices. When chosen, the chart or graphic comes into view. Web page constructors (Wix, Weebly and Adobe Dreamweaver) also have this capability. You can discover the specifics after some exploration of the display choice options in each software. The main idea is to keep the surrounding information static while the chart or graphic changes. This could be a hyperlinked image or a slide transition that 'switches' only the image of the chart or graphic that corresponds to the user choice from a menu. The menu could be a dropdown or simple radio buttons. Tableau, Excel and Mathematica also have this capability, with the added ability to recalculate the data. An image is static and once recorded, will not change. Dynamic can also mean the ability to be updated. PowerPoint can have dynamic links to an Excel chart so the data can be updated. Tableau has an internal timer that can be set to automatically update online data.

Creating infographics

Chapter 4 outlines what an infographic looks like. This section takes a step-by-step tour of how to create your infographic. You will be referring to your preferred design mock-up from Chapter 4 (which was submitted as part of Unit 3, Outcome 2). This will be the basis for your infographic construction. Remember to note any changes you make on your Gantt chart. Be sure to take notes, so you can later annotate your final Gantt chart to show the reasons why changes were made.

Ad The NelsonNet resources have examples of Excel, Tableau and Mathematica dynamic data visualisations.

A sequence of slides shows how user choice can change the view. Slide charts and graphics can be made to appear to change, when it is actually the slide that is changing. The chart is changed by a button hyperlinked to the slide with the next chart.



SCHOOL-ASSESSED TASK TRACKER

☑ Project plan	☑ Collect complex data sets	☑ Analysis	☑ Folio of alternative designs	☑ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

After choosing your presentation software, there are a few common steps that apply to every project. Whether you choose an infographic or a dynamic data visualisation, there are a few checklist items that apply to both formats.

1. Keep it simple. The purpose of the infographic is to communicate a message, not to use a large number of visual effects, colours or fonts.
2. Check your spelling and grammar as you go. Use the spell and grammar checker that is built-in to Word, Pages and LibreOffice.
3. Use a grid. Grids are available in a variety of shapes and sizes. You will probably use a two, three or four-column grid that will help you arrange text and images (Figure 5.29). More columns give more flexibility to allow for different sized text boxes and images and retain alignment.
4. Choose colours carefully. Ensure your choice of colours supports your message, rather than distracts.
5. Balance the amount of text, graphics and charts. Consider the 'flow' as the user reads your graphic solution.
6. Ensure interactive elements are obvious. If you have a dropdown menu or hyperlink, be sure it is apparent and not seen as emphasised text. Perhaps use an icon to indicate a user control.
7. Are all elements included? Title, research question, findings statements, charts, images, graphics, reference list of sources other acknowledgements and your details are some elements you may choose to display within your infographic.
8. Is navigation obvious? If there is more than one page, how does the user navigate?

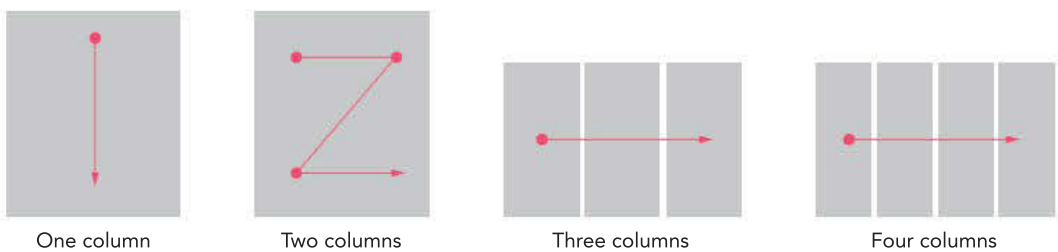


FIGURE 5.27 Typical grid template with one, two, three and four columns when reading from left to right, top to bottom

Creating dynamic data visualisations

The development of an infographic may be thought of as operating on one layer to produce a story with charts to illustrate a message. The development of a data visualisation can be considered a window into several layers of information. The added capability of adjusting the view by changing user controls allows for greater storytelling capacity. Your main task is to assemble the elements to allow a clear message to be presented.

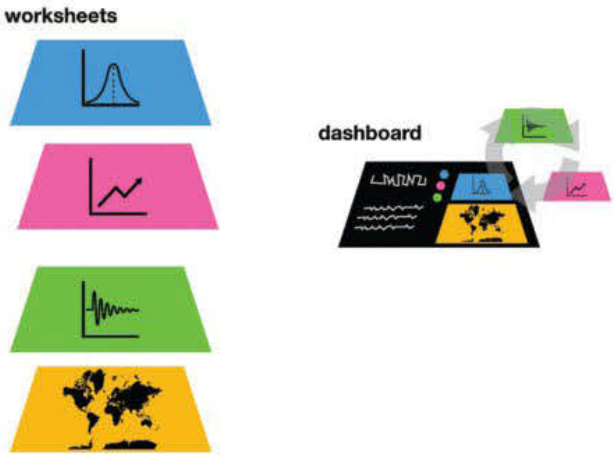


FIGURE 5.28 A dashboard can be considered a summary of several layers or worksheets. Radio buttons allow a chart view to be changed. The dashboard could be in PowerPoint, and the worksheets in Excel.

Macros in Excel – a warning!
Many of the features of Excel, while suited to the development of the Unit 4, Outcome 1 dynamic data visualisation, require a disproportionate amount of time to learn the details. Macros, for example, offer convenient methods of quickly adjusting the appearance of a chart or navigating through several worksheets. Macros can be written or recorded; however, the set of instructions or code to execute the action may be more easily achieved by simpler means.

Form tools in Excel

Excel offers several tools to include interactivity to your presentation. Hyperlinks could offer a similar functionality in PowerPoint. Various controls available are:

- buttons
- checkboxes
- scrollbars
- option button (also called a radio button).

Before you can use Excel Form Controls, the Developer tab needs to be enabled.

To enable the Developer tab:

- In macOS, open Preferences > Ribbon & Toolbar (Figure 5.29a).
- In Windows, go to File > Options > Customize Ribbon (Figure 5.29b).
- Under the Customise the Ribbon box on the right, you will see all available tabs.
- Select the checkbox next to the Developer tab.
- Save your choice.

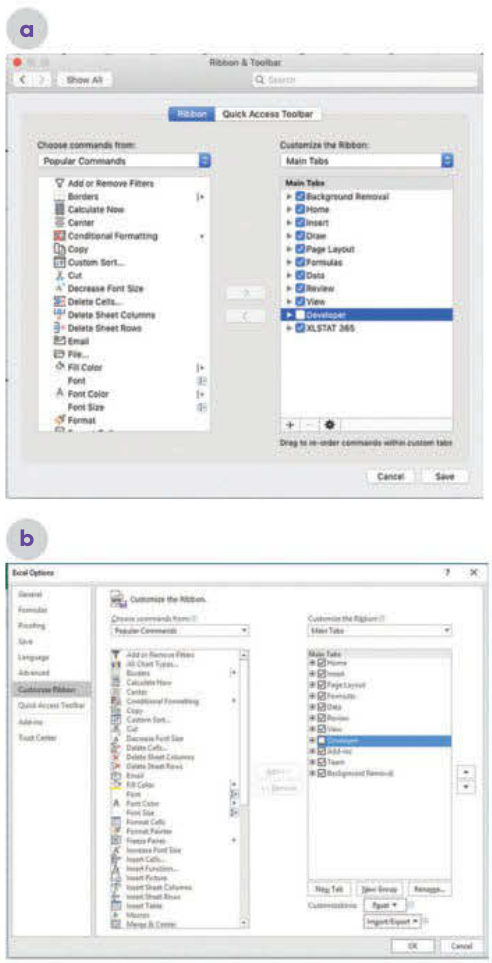


FIGURE 5.29 Changing Excel preferences by adding Developer tools for **a** macOS, and **b** Windows 10

SCHOOL-ASSESSED TASK TRACKER

☑ Project plan	☑ Collect complex data sets	☑ Analysis	☑ Folio of alternative designs	☑ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

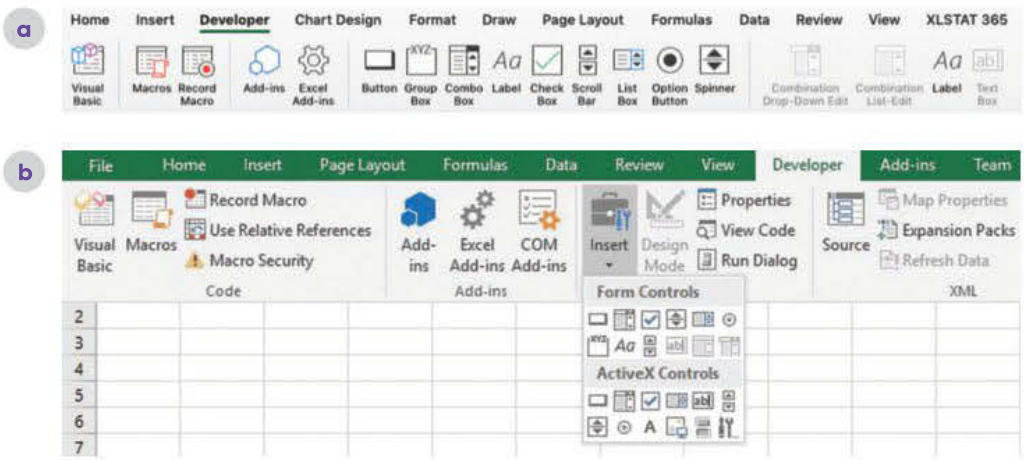


FIGURE 5.30 Form Tools in Excel allows action controls to be inserted into a worksheet or dashboard to allow users to be empowered to switch from one view to another with a list of simple choices: **a** macOS, and **b** Windows.

How to use Form Controls in Excel
This Microsoft tutorial contains step-by-step directions on how to include form tools in your worksheet. It is beyond the scope of this text to provide that level of detail.

Visual Basic

Excel also has the capacity to respond to Visual Basic programming. This behind-the-worksheet coding will also achieve interactivity; however, the skills necessary for this tool are beyond the requirements for Data Analytics. (**Note:** There is no requirement for coding skills in Data Analytics.)

(Windows) Notepad++
Animatron
Audacity
Windows Movie Maker
Inkscape
GIMP

Multimedia authoring

Even the most modest of smartphones now offer facilities to create and edit documents incorporating images, video, audio, animations and text, though you should still approach multimedia authoring with care and thought.

Media editors that may assist you with your solution are listed in Table 5.6.

TABLE 5.6 Media editors to assist with multimedia authoring

Media editor	Description
Image editors	Adobe Photoshop (expensive, powerful), GIMP (free), Inkscape (free), Keynote (free), Preview (free, very powerful, macOS only) Enables users to crop, convert to other formats, resize, distort, repair, combine images, add text and shapes, and add special effects to images
Video editors	Windows Movie Maker (free), Avidemux, Videoredo, AVS video converter, Adobe Premiere Pro (very complex, powerful and expensive), iMovie (free, macOS), Clips (free, iOS only) Can read video files in several formats and join or cut scenes; add soundtracks; add special effects like credits, fades and dissolves; distort colours; and export to different formats
Audio editors	Audacity (free), Adobe Audition (expensive, powerful), GarageBand (free, very powerful, macOS and iOS) Can crop, cut and join clips; repair clicks, hums and background noise; change volume; merge and mix multiple tracks; save in a variety of formats
Animation editors	Animatron (free, online), Keynote (free, very powerful), Adobe Edge (powerful, expensive) Create a timeline in which users can add and edit video, audio and text to create engaging videos
Text editors	Microsoft Word, WordPerfect, Microsoft Publisher, PowerPoint, Notepad, Pages (free, macOS and iOS), Keynote (free, macOS and iOS and on iCloud for winOS) Fundamentally important, whether they are high-end, like Microsoft Word, or a basic text editor, like Notepad; advanced text editors, such as Edit Pad and Notepad++, offer multiple tabs, regular expressions, macros and code colouring for programmers

9780170440875

Verification and validation

Verification is the process of checking that the final product software meets the requirements detailed in the *design* stage. The process of verification forms part of the evaluation of the project. The evaluation criteria are considered and as developer of the graphic solution, you decide when the criteria have been satisfied. A formal evaluation will take place in the *evaluation* stage later in this chapter. Validation, however, is entirely different and ensures the data is appropriate.

Validation techniques

Validation checks that the input data is *reasonable*. Validation does not and cannot check that inputs are accurate. How could validation tell whether a person is being honest when entering their age? Validation can detect problems when a person enters their age as 183 years, or 'apple', or nothing at all. You can perform validation manually (yourself) or allow software to do it for you.

Computers are particularly good at conducting validation checks.

1. **Existence checks** ensure that a value has been entered and the field is not blank or <null>.
2. **Type checks** ensure data is of the right type; for example, that the age entered is actually a number.
3. **Range checks** ensure that data is within acceptable limits (for example, children enrolling in kindergarten must be 3–6 years old) or comes from a list of acceptable values (for example, small, medium or large).

People can perform manual validation, especially proofreading for sense, clarity, relevance and appropriateness. Values entered into a spreadsheet may pass electronic validation checks but can be inaccurate because they are ridiculous. Microsoft Word will point out words not in its dictionary but it cannot advise if an extract is misleading or factually incorrect. A person will smell a rat where a computer cannot!

Remember:

- **Validation** checks the reasonableness of data inputs.
- **Testing** checks the accuracy of information outputs.

In your SAT, ensure that all of your data is thoroughly and appropriately validated.

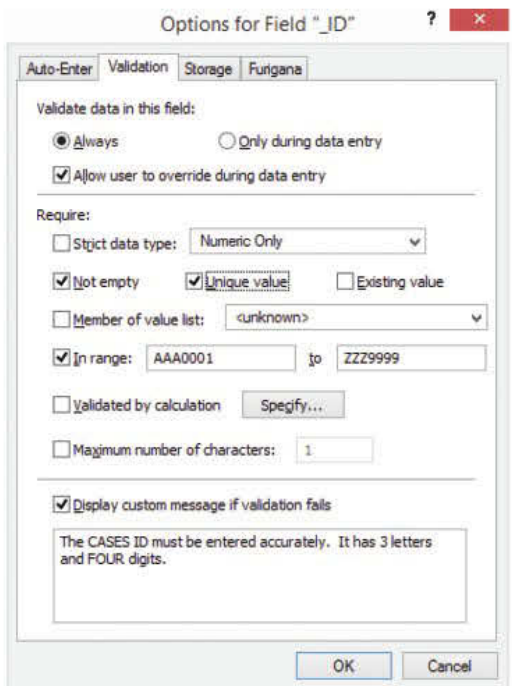


FIGURE 5.31 Validation rules in FileMaker database. Here, an ID field is made compulsory ('Not empty') and unique, and within a defined range of values. The database is also told what error message to display if validation fails.

THINK ABOUT DATA ANALYTICS 5.6

If the person is expected to enter their age, what would be a reasonable range check?

a 5–50 years
b 15–80 years
c 0–100 years
d 1–200 years?

THINK ABOUT DATA ANALYTICS 5.7

In Unit 4, Outcome 1, which data will you need to validate, and how will you achieve this?

9780170440875

Testing

If a software solution fails, it could annoy or disadvantage users, so thorough and careful testing is necessary whether the solution is a game, a website shopping cart, or an airliner's autopilot.

If your solution fails because of undiscovered faults, it may become difficult to use, or be completely unreadable.

Testing checks that a solution produces the correct output and does what it should do. Testing is not easy, quick or cheap – especially for a product such as an operating system with megabytes of code in thousands of files created by hundreds of people.

The typical steps involved in testing are as follows.

1. Decide which tests will be conducted.
2. Create suitable test data.
3. Determine expected results.
4. Conduct the test.
5. Record the results.
6. Correct any errors.

There are many testing types, each intended to uncover different kinds of errors at different times during development. The types of testing relevant to your solution are listed in Table 5.7.

TABLE 5.7 Testing types

Type	What is tested?
Informal (alpha)	This is the part of the solution that has just been finished.
User acceptance (beta)	Typical end-users use their own equipment to check that the finished solution is acceptable under different user conditions.
Component	A single part of a system works properly by itself (for example, a user entry form applies the correct delivery cost for a given destination).
Integration	Individual parts of a system work together (for example, the embedded Excel file correctly accesses the separate database table).
System	All components in the solution work properly as a single unit.
Installation	The form control is installed correctly and working on your system, server or domain.
Compatibility	The multimedia plug-in and its components are compatible with a variety of browsers and the main OS. **Note:** ActiveX Form Controls will not work on the macOS.
Usability	This is whether users can operate your graphic solution quickly and simply.
Accessibility	This is whether users with additional needs or disabilities can use your graphic solution.

Test data

To prove the accuracy of a solution's output, give it some test data to work on, and compare the solution's answer with one known to be correct.

Good test data includes the following.

- Valid data – data that is perfectly acceptable, reasonable and fit to be processed.
- Valid but unusual data – data that should not be rejected even though it seems odd. A 12-year-old genius might want to enrol in a VCE subject.

9780170440875

- Invalid data – to test the code's validation routines. For example, if people must be at least over 60-years-old to obtain a Senior's Card, test data should include people *under* 60 so they can be seen to be rejected.
- Boundary condition data – data that is on the borderline of some critical value where the behaviour of the code should change. These 'tipping point' errors are a frequent cause of logical errors in programming.

Testing your solution

After designing and building your infographic or dynamic data visualisation solution, you need to demonstrate that it has been thoroughly tested. You need to know what to test in your solution. The following sections discuss these.

Media and plug-ins

You must inspect each image, audio clip, video, graph and animation (that is, any non-textual information) to confirm that it is displaying in the right place, at the right time, at the correct speed and volume, in a variety of common environments; that is, in different browsers and devices.

Hyperlinks

Manually click on every internal and external link in the solution and note the result. Create a list of links and tick off each one as it passes testing.

Links to external services

You should be able to test all parts of the solution under your control completely. You need to test the operation of any external connections to your product to ensure that data updates function as expected.

Readability

Use the checklist provided in Table 5.8 to test the readability of your solution. If your solution does not meet requirements, you may need to rethink your design.

TABLE 5.8 Readability checklist

Checklist	Tick ✓
Is the text large enough to read comfortably on a small device?	
Is contrast optimal, or at least satisfactory?	
Is the typeface a readable size?	
Are lines or paragraphs of a good length?	
Is text alignment attractive and readable on the page?	
Are the spelling, punctuation and grammar correct?	
Is the vocabulary appropriate and inoffensive?	
Is expression clear and unambiguous?	
Are headings clear, and do they divide content into logical sections?	
Are all charts appropriately labelled?	

9780170440875

Calculations

If your solution calculates any information, its answers need to be verified by manual recalculation in a testing table. For example, you might create a web page containing JavaScript code to display a countdown timer until the next Data Analytics examination. To prove that you have tested the accuracy of its output, take a screenshot. Annotate the screenshot with whatever manual calculations will demonstrate that it is true, based on the time the screenshot was taken and the time of the exam.

Loading times

If the graphic solution is online, clear your browser's cache to remove pre-loaded copies of files and media and try loading the site via cable and wi-fi. Any page that takes more than a few seconds to load should be inspected and optimised. Another method is to use one of many online services that can measure the loading times for your pages. Online data repositories may have varying access times due to user demand.

Browser compatibility

Does your solution rely on a browser for presentation? Check that plug-ins and installed players and codecs (coder/decoder or compressor/decompressor) can read and display your chosen media. Browsers differ in their ability to interpret different media, and some systems may not have the right technology, such as HTML5, or the necessary plug-ins installed, such as Flash. Every piece of media must be checked on the dominant browsers to verify that they appear as expected. Remember Flash will not play at all on many mobile devices.

You can test most aspects of site functionality yourself manually, but if your solution is online, there are many services available that can perform automated cross-browser compatibility checks using new and previous browser versions.

Accessibility

Does your solution create unnecessary difficulty for users with poor eyesight or muscular control, weak language skills, or other common disabilities? Is alt text applied to images? Are colour combinations considerate of people with colour vision deficiencies (CVD)? Many CVD considerate palettes are documented online.

There are several places online to test the accessibility of your solution – try the World Wide Web Consortium (W3C) web page for one such website.

Dynamic features

Every selection option item must be checked and its behaviour documented in a testing table (Table 5.9). If data entry forms are expected to work, data should be entered and its successful arrival at its destination should be documented. Any simulated functionality, such as a faked login box, should, as far as is practical, appear to work genuinely. Any coding, such as JavaScript, PHP/MySQL, Java, Perl, macros and Python, should be run using a variety of test data, and the behaviour of the code recorded.

Testing table

A **testing table** is a common method to record evidence of functionality testing. A testing table for a data visualisation may look like Table 5.9 (page 247).

9780170440875

TABLE 5.9 A testing table

What was tested	How it was tested	Expected result	Actual result	How it was fixed (if relevant)
Embedded YouTube video	Loaded page and clicked play button.	Plays cat video smoothly.	✓	N/A
Links to pages for contacts.html, references.html, conclusions.html, and data.html	Clicked each link on 'index.html' (links on other pages were copied from here, so if they work on the homepage, they will work everywhere else).	Go to linked pages.	All worked except for data.html.	Changed link text in source code from 'data.htm' to 'data.html'.
Readability	Asked two volunteers to read sample pages and report on text size, contrast, alignment, spelling, vocabulary, offensiveness, and headings.	Reports that pages were easy to read, accurate and inoffensive.	One reader suggested a paragraph in italics was hard to read.	Changed paragraph's text style from italics to bold.
JavaScript calculation of number of days since the page was created	Set computer's clock to 7 days after the page creation date.	Index page should display 'This page was created 7 days ago, so some external links may be no longer accurate.'	Displayed '... created 6.89586 days ago ...'.	Rounded up the age calculation in JavaScript.
Loading times from databin	Performed website speed test on each page.	Performance grade of at least 75	All pages were graded 77/100 – 'faster than 87% of all tested websites'.	N/A
Browser compatibility	Load all pages in: • Chrome, Firefox and IE/Edge (all on PC and a tablet) • Safari (on iPad)	All pages appear as intended in the site's design.	• Chrome v.41: all pages good • Firefox v.35: all pages good • IE v.11: index.html containers not aligned as they should be • Edge: all pages good • Safari: pages okay, but controls too small	• IE: had to change the spacing of elements; this didn't upset the other browsers during re-testing. • Safari: increased control sizes 20% to make them more finger-sized.
Alt text	Used Windows Narrator to read the data visualisation text. Alternatively, use macOS VoiceOver to read the text.	Alt text should be read by Narrator/VoiceOver.	All alt text was read okay.	N/A

9780170440875

Classroom constraints

Your dynamic graphic solution may, because of constraints in classrooms and networks, not have access to updated data online. However, it should have the look, feel and apparent functionality of a real online solution, even though some features may have to be simulated because it is unreasonable to expect them to function under all working conditions.

How to document your testing

The following is more or less a checklist of how to approach documenting your testing.

- Use a testing table (such as the one shown in Table 5.9).
- Seek a subjective report from a fellow student who tried out your solution's readability and usability.
- Capture screenshots of features that are not normally visible, such as dropdown menus and warning messages, showing that they work when needed.
- Make handwritten calculations annotating printouts of screenshots of your solution's calculations to verify that the output has been checked for accuracy.
- Capture screenshots of the solution's validation rules responding properly to invalid data.

Evaluating your solution

Evaluation is not the same as testing; its purpose is distinctly different.

Evaluation is the final stage of the problem-solving methodology. It checks how well the solution is satisfying the needs of the user for which it was originally created.

Remember: You will be evaluating your dynamic data visualisation solution and your project plan as part of your assessment in Unit 4, Outcome 1. When evaluating your solution, you need to refer to the evaluation criteria you developed during the design phase (page 185). For each criterion, you will choose a method to use for its evaluation.

Evaluation is not the same as testing; its purpose is distinctly different. By the time evaluation begins, the solution has already been proven to work properly and its functionality is no longer in question.

Evaluation can best be understood by saying what it does not do. Consider the following.

- Evaluation does *not* test that a solution is working properly. That should have been done during testing.
- Evaluation does *not* enter test data to check that output is accurate. That should have been done during testing. Comparison between one version and another version of test results is evaluation, however.
- Evaluation does *not* use a stopwatch to time how long a process takes. That should have been done during testing. However, comparisons of test times would be legitimate evaluation, referring to test results.
- Evaluation does *not* perform checks with immediate results, such as pulling out the power plug to see if a system loses data. That should have been done during testing. Evaluation looks at a solution's performance *over time* in terms of the **evaluation criteria**.

What to evaluate

Evaluation criteria are determined during the design phase of the problem-solving methodology and are based on the most important qualities that the solution is expected to have when it is designed. For example, for your solution, essential criteria include appeal to your intended

9780170440875

audience and adequate substantiation of the conclusion of the results of your research question/topic. For an accounting package, security and accuracy might be most important. You should evaluate the features that would, if they were not achieved, render your solution unsatisfactory.

Evaluation criteria fall under two headings: efficiency and effectiveness, as described here.

1. *Efficiency* can be measured in terms of speed or productivity (work produced in a given time), profitability (income generated versus running costs) and labour requirements (how much labour is required to achieve its productivity levels).
2. *Effectiveness* includes completeness, readability, attractiveness, clarity, accuracy, accessibility, timeliness, communication of message, relevance and usability (see Figure 5.13, page 221).

Conduct interviews or create questionnaires to seek feedback from your peers that is relevant to the criteria you expect your solution to achieve (such as communication of message, readability, accuracy, clarity and so on). This feedback could be offered as evidence of evaluation.

For a dynamic data visualisation, you may decide to use annotated screenshots to demonstrate how key criteria were satisfied, such as readability or consistency of formatting. If recordings of animated screen activity are needed to prove you satisfied an evaluation criterion, you could use screen recording software such as Camtasia. To demonstrate that you achieved appropriate readability levels, you could include a screenshot of your text's readability score, as calculated by Microsoft Word (Figure 5.16 shows an example of this).

Evaluation methods

For each evaluation criterion for your solution, there must be a corresponding evaluation method that can measure the degree to which the criterion has been achieved.

- **Objective** (fact-based, measurable) results are solid facts that are hard to argue with. Measure whenever you can.
- **Subjective** results (emotions, opinions, personal judgements) can be gained from interviews, questionnaires and surveys. These should only be used when objective measurement is not possible or practical, such as to evaluate how comfortable users feel when using a multimodal solution.

TABLE 5.10 Typical evaluation methods

Criterion	Method
Accuracy (effectiveness)	Check the complaints log and count the complaints from staff or customers about inaccurate information received over three months from the system.
Reliability (effectiveness)	Count the number of faults in the system's error log.
Security (effectiveness)	Count the number of successful and thwarted attempts made to penetrate system security.
Attractiveness, pleasure, comfort, confidence (effectiveness)	Interview users.
Productivity (efficiency)	Refer to system logs and count how many transactions the system handled over three months compared to the previous system.

9780170440875

Criterion	Method
Profitability (efficiency)	Ask the accountants to tally the new system's running costs over time. Check organisational profit figures and see if it has increased.
Labour requirements (efficiency)	Count the number of staff hours spent operating and maintaining the system compared with the previous system.
Ease of use, usability (effectiveness)	Count the number of times online help was used (indicating that the solution may not have been intuitive). Add up how many errors were made by users. (A solution that is hard to use tends to cause users to make mistakes.) Check the help desk records to see how often users asked for help or complained about the solution. Give users a questionnaire asking about their feelings regarding the system's usability.

Remember: Evaluation assesses your solution's performance over time. It is not instantaneous the way that testing is. Any emotional or judgemental feedback is only gathered on appropriate criteria. For example, it is pointless to ask interviewees questions such as, 'Is the new system faster than the old one?' Even if you received an answer to this question, you would not be able to trust its accuracy.

When to evaluate

Evaluation occurs after the solution has been in regular use for some time so it is well 'bedded in' and its users are familiar and comfortable with it. A few months of regular, daily use is typical.

Evaluating a solution too soon can lead to negative feedback if users are not yet used to it and are slow and prone to making errors. Later, when they are comfortable and skilled with the solution, their feedback may be much more positive.

In cases when a system is used infrequently, but its success is critical to the organisation (such as creating school reports, or managing a flood of tax returns at the end of the financial year), evaluation may be done immediately after the system is used.

Waiting is not a feasible option for the solution for your Outcome. You will not have the luxury of waiting for a long period of time for users to explore your solution. Instead, you will need to ensure that you evaluate your own solution thoroughly and take into account all of the feedback you received during beta testing, which should have included asking potential end-users to look at your solution. (Potential end-users could include classmates, friends, your teachers, your parents and other family members.)

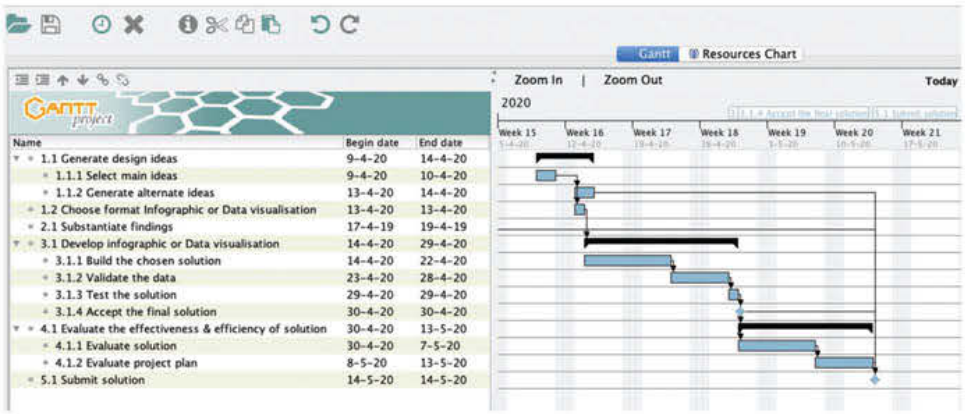
Documenting the progress of projects

In the first half of the Data Analytics SAT (Unit 3, Outcome 2), we discussed how to manage a project, and why project management is important. In Unit 4, you will continue to use these project management techniques to complete your research by creating an infographic or dynamic data visualisation solution.

9780170440875

It is unlikely that any project will ever proceed perfectly in line with the project plan. One unexpected rainy day, a hard disk crash or a sick day for a key worker can slow down a work team enough to affect tasks, other teams and deadlines. Project plans are not written in stone and obeyed regardless of real-world events. Gantt charts should be modified regularly to reflect reality.

Project plans are living documents. Tasks that run overtime may have resources added to them or be modified so they finish earlier. Bad weather may force changes to scheduling, so indoor tasks may need to be attended to instead of outdoor work. Late deliveries of equipment may cause a project manager to take people from one task and assign them to another that can proceed without the deliveries.



Created using GanttProject, https://www.ganttproject.biz/

FIGURE 5.32 Gantt chart: a project plan for the chosen solution

NelsonNet additional resource: Figure 5.32 Sample GanttProject chart.

Gantt chart

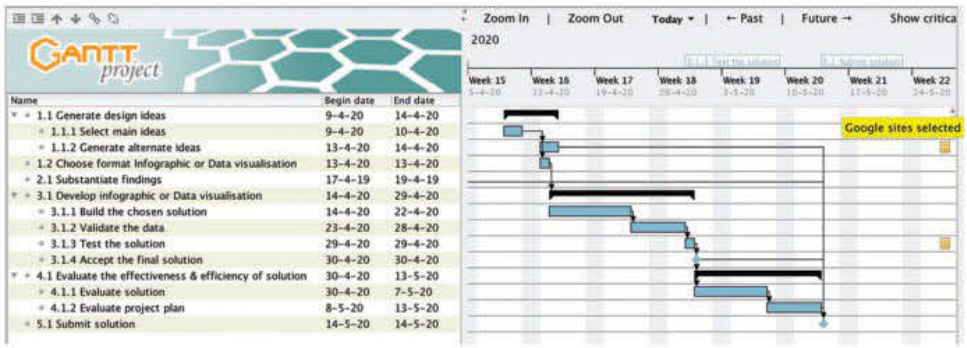
RESEARCH

1. Redraw Figure 5.32 by hand or electronically, or modify the Gantt chart in Chapter 4 of the student website (Figure 5.32 'Additional resource' icon linked here) to add two days to the '2.1 Substantiate findings' stage.
2. Adjust the rest of the project so that the final date does not move forward by more than one day.
3. Justify why you have chosen to reduce the length of at least one stage by one day.

A project plan may be annotated to explain reasons for changes to task schedules or resourcing priorities. When the project is later assessed, these annotations might serve as valuable lessons when undertaking the next project. Annotations may also be added by other project leaders to advise the team of significant news or concerns. You could annotate by hand, or make notes in the Gantt chart itself (Figure 5.33, page 252).

Project logs are a record of all the small and large steps a project takes on its way to completion. It is usually in electronic form and shared with all project leaders. It may be created with specialist log software in a weblog for example – or with Microsoft Word or Excel. You could share it online using Google Drive, or with a similar technology.

9780170440875



Created using GanttProject, https://www.ganttproject.biz/

FIGURE 5.33 Gantt chart with notes added

Figure 5.34 shows an example of a project log composed in Microsoft Excel. Columns C, F, G, H and J use data validation dropdown lists. Columns G and H also have conditional formatting applied to indicate priority and impact by colour: low, medium, high, critical and showstopper. The template for this project log is available on the NelsonNet student website.

Issue #	Date	Reported by	Item/Functionality	Description	Issue type	Priority	Impact	Suggested fix	Status
1	2020-05-09	Developer	Graph in data visualisation	Interactive sequence is out of order	Functionality	High	High	Check the sequence order, and name of graphic	Resolved, incorrect naming of graphs
2	2020-05-09	Developer	Title	The font size of the research question is too small, hard to read.	Design	High	High	Adjust the size and colour contrast	Resolved, larger font, darker colour
3	2020-05-09	Developer	What to do is not clear	After reading the title and research question, the interactive buttons are not clearly marked.	Usability	Medium	Medium	Adjust the size and colour shading of the button. Add a label	Partly resolved, need to work out how to include a label inside
4	2020-05-09	Developer	Image quality	Image three is very poor quality	Data	Medium	Low	Locate better original and re-capture better quality screenshot	Still looking to locate image
5	2020-05-10	Reviewer	Sources	There is no acknowledgement of sources	Data	Low	Low	Add Sources citation to lower left of page	Yet to be completed

FIGURE 5.34 Project log template created in Microsoft Excel

NelsonNet additional resource: Figure 5.34 Project log template.

A project log helps a team to keep track of project tasks, such as:

- which team member is responsible for the tasks
- when deadlines and milestones are due
- the status of tasks.

It can also help you to keep track of the work you need to do on your solution yourself. In conjunction with your Gantt chart, it can be a valuable tool to help you manage your project efficiently.

A project log can include:

- time and date stamps
- comments on progress
- project risks
- issues that arise
- ideas for solutions
- explanations of decisions
- results of testing
- forecasts and warnings
- task changes
- photos
- future action needed.

9780170440875

A project log is like a diary that describes the complete history of a project. An online version would be to keep a **weblog** or **blog**. Keeping a weekly blog that records daily achievements would be a very effective method of gathering evidence. Your final Gantt chart can be annotated with entries from your blog, with references to the URL.

If you find yourself struggling to divide up the tasks by priority, creating a priorities quadrant may also be of use (Figure 5.35).

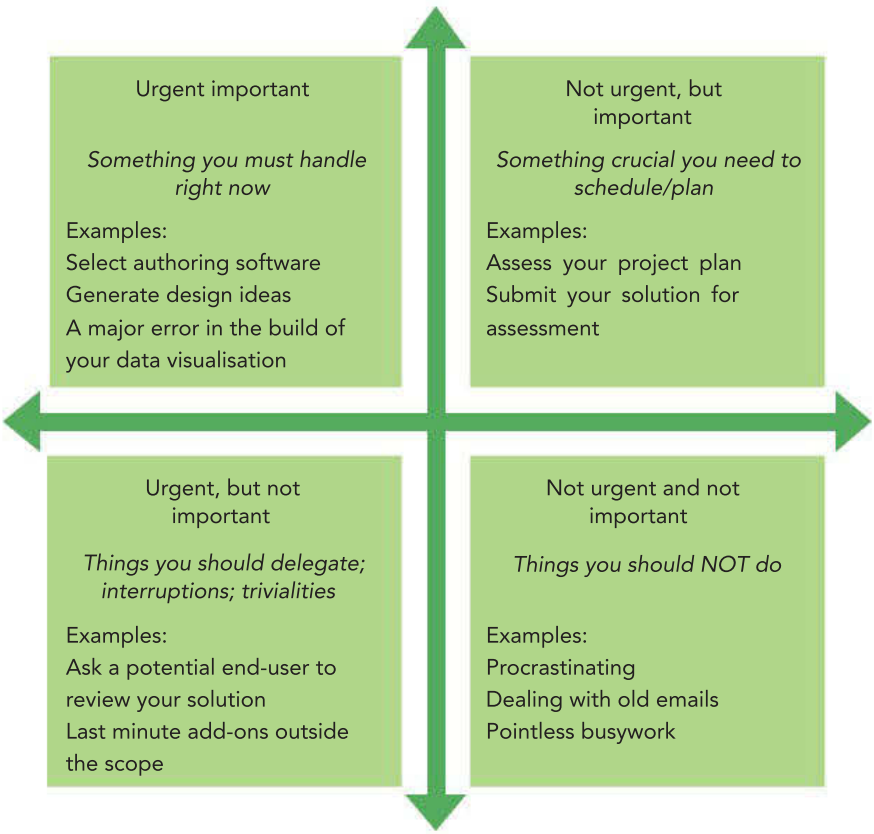


FIGURE 5.35 Example of a priorities quadrant

Efficiency tips

You have a limited amount of time to complete the project, so you need to make sure that you use your time as efficiently as possible. Your first priority is to make the solution that you are proposing work. While the attractiveness of your solution is an assessment criterion, appearance is worth much less than the quality of its information. You need to make sure that you allocate your time appropriately to reflect this priority.

One way of doing this is to use a plain text heading and then insert all of the key information. If you have time later, design a graphic heading and modify the choices of fonts. Remember, a beautiful solution with incomplete content is a demonstration of wasted time.

It is easy to waste time 'fiddling' with small details. It is not an efficient use of of time to spend half an hour in Photoshop cutting a figure from its background when there are more important things to be done. You need to complete the less exciting jobs, such as typing data and testing that all links work.

9780170440875

It may help to make an agreement with yourself to do just a little of the task at a time – in this way, you will at least make incremental progress. Your self-confidence will be boosted as you complete small portions of the work, and you may even decide that since you have started you might do a little bit more while you are at it. You will find once the first step has been taken, it is usually much easier to continue.

Assessing your project plan

Organisations invest a great deal of time, money and labour into projects, so they tend to look back at their project plans to evaluate how project planning went. Organisations often need to undertake further projects, so they need to keep lessons in mind from earlier projects to prevent repeating mistakes. Evaluating a project plan can help answer questions like the following.

- Did the project finish on time?
- What tasks delayed the project? Why were these delays not anticipated?
- Could lessons be learnt to help the next project finish on time?
- Did the project finish on budget?
- What assumptions were wrong?
- Why did this task cost far more than expected? How can this be avoided next time?
- Why were new requirements being added just weeks before the system was due to go online? Was our analysis a failure?
- Why did the first three prototypes blow up? Was the design team under-skilled, overworked, under-equipped or working to an impossible deadline?

THINK ABOUT DATA ANALYTICS 5.8

Identify the questions from the list opposite that evaluate an organisational project plan that could also be used to evaluate your project plan.

Even a failed project can be a valuable learning experience if it prevents the same mistakes happening again.

On the other hand, it can also be an expensive mistake to *not* undertake projects to improve systems. Below is an example of this situation.

> As of 2012, Long Beach, California was owed $18 million in parking fines as a result of antiquated software used by the local government. Fines had gone uncollected because of the age of the systems in use. Staff were stuck using manual processes that took up so much time that there was no time left to undertake collection efforts – leading to $18 million in unpaid fines. Even worse: Long Beach first knew about the problem as of an audit in 2009, but still failed to address it (*LA Times*).

Your project may be on a more modest scale, but there are lessons that you can learn. The purpose of assessing your project plan is to judge how your plan, and the techniques you used to make adjustments along the way, helped you to manage your project. For instance, how much notice did you ultimately take of your plans? How did your plan assist you when things did not go as expected? Did the quality of your annotations or other tools you may have used keep you on track? These and other questions are of great use to you now, and they are a good idea to keep in mind for your future studies and later career, because they will help you to understand how planning functions alongside real-world projects.

You need to fully think through the scope of the tasks so you do not discover a forgotten task when your time has nearly elapsed and it is too late to finish it. Make sure your work breakdown structure (WBS) is complete. Keep referring to your project plan so you do not forget tasks, do them in the wrong order, or fail to observe milestones.

9780170440875

You also need to think carefully about your software choices. It would be unwise to choose a visualisation authoring tool, only to discover halfway through development that it is unsuitable for the task. If you need to learn and practise new software skills, do it before the project begins. Learning new software while you work on a critical solution is inefficient and you will be prone to making errors.

You also need to be realistic when you are estimating time requirements. A good thing to do is think back to when you had similar tasks to complete. Remember which problems caused delays and add time to your schedule to deal with them this time. If a task looks like it will run overtime, refer to your Gantt chart to determine the impact it might have on the project. If the task has slack built into it already, it may have no effect on the tasks that follow. If it is a critical task, you will need to save time on other tasks, and find where and how this can be done.

When evaluating your project plan, you first need to establish the evaluation criteria that will indicate how successful it was in managing your activities and getting the project completed on time.

Such criteria may include the following.

- Completeness: Were any significant tasks omitted from the WBS? Were resources included? Was it annotated when required?
- Maintainability: How easy was it to modify the Gantt chart to keep it up-to-date with reality?
- Accuracy: Were tasks correctly identified and marked as dependent or concurrent? Were tasks in the right sequence? Were time estimates realistic?
- Readability: Was it easy to see all tasks and their dependencies? Was the chart and its text of a readable size? Were colour choices appropriate?

Once complete, and relevant criteria have been chosen, you will again need a method to evaluate each one. Annotated printouts highlighting key features of the solutions may be useful. You might take screenshots before and after changing a task's duration to show how easy it was to maintain the chart. You can describe how well the chart worked to keep you on schedule to finish the solutions on time. You might offer lessons you learnt from the projects that will make later projects even more successful.

Next steps

Work your way through the chapter summary material, including 'Preparing for Unit 4, Outcome 1' and, following your teacher's instructions, prepare your solution for submission.

9780170440875

5 CHAPTER SUMMARY

Essential terms

business intelligence (BI) actionable knowledge attained from dashboard summaries of manipulated and analysed raw data

comprehensive management plan a file management strategy that includes storage, retrieval, backups, archiving and security of documents

condescending showing an attitude of superiority

convention the accepted way that meets expectations when a particular format is used

data visualisation the presentation of data in a pictorial or graphical format

disaster recovery plan (DRP) details of the steps required to recover information systems in the event of damage or loss

dynamic data visualisation a graphic presentation of data that can be altered in order to change the relationships between data or aid understanding

dynamic versatile disk (DVD) compact disc that can typically store 4 GB or 8 GB with two layers

evaluation assessing whether a solution achieves the goals for which it was originally designed; not the same as testing

evaluation criteria rules set out during design that include effectiveness and efficiency criteria; based on the solution's requirements that were defined during analysis

existence check validation that ensures a field has data entered into it

format how something is displayed (for example, using a table or graph); arranging the look or presentation of an object

hard disk drive (HDD) a magnetic platter with a read/write mechanism

mandatory convention a convention required by law or by-laws, which must be adhered to without exception

mean time between failures (MTBF) a statistical estimate of the expected reliability of a computer component; usually read/write actions for drives

null field a field that is empty when a value is expect, which can generate an error in the manipulation software

objective fact-based, measurable; a type of study that requires the participant to remain detached from the activity and apply previously agreed behaviours, rules and protocols

optional convention a convention where a degree of choice is available, and the rules may or may not necessarily be followed

plug-in a piece of software that adds onto an existing computer program

preferred convention a convention where some choice is available, though some are more preferred (or usual) than others

range check validation rule to ensure that data falls within an acceptable limit

readability the ease with which the reader can understand a specific text

relevance staying on topic

solid state drive (SSD) a disk drive with no moving parts; SSDs have an estimated 'lifetime' dependent on the number of read/write actions

subjective based on opinion, emotion or personal judgement; for example, 'Is that salesperson friendly?' (the opposite of **objective**)

9780170440875

testing ensuring that something works as intended and checking the accuracy of information outputs; that they are correct with no errors, or that any errors can be dealt with without inconveniencing the user

testing table a table set up to record functionality testing (what will be tested, how it will be tested, and what the expected result will be if the elements work as expected)

type check validation rule that checks that only the correct data type is accepted

validation checking that data input is reasonable, which is an activity within the development stage of the PSM; typically used for existence, range and type checking

verification the process of checking that the final product software meets the requirements detailed in the design stage of the PSM

version control management of files and software so only one version of the most up-to-date file is in existence at a time and 519 so that major reworkings can be reversed if needed

weblog (or **blog**) an online journal or diary with date and timestamps; files and images can be uploaded as evidence of progress

wiped when data is scrubbed and overwritten with 0s and 1s

Important facts

1. **Dynamic solutions** may use more than one type of data, such as text, images (still or moving), audio and animation. The viewer can choose or specify aspects of the displayed solution. In some instances, the source data can be updated at specified intervals, or continuously if online.
2. A successful **solution** must convey its message clearly and be easy to understand.
3. Your graphic solution (infographic or data visualisation) must use certain types of software.
4. To improve **inclusive language**, use culturally neutral vocabulary and references in your solution that may be understood by other cultures. Use universally understood terms and units of measurement, such as date formats. Aim for clear, concise language that non-native English speakers can easily understand.
5. A **target audience** may include readers of various ages. Try to write in a way that is accessible to as many of them as possible.
6. **Formats** are chosen methods of presenting information (for example, tables, graphs, web pages).
7. **Conventions** are the standard ways of using formats (for example, underlined weblinks) that facilitate information absorption.
8. **Design principles** affect the functionality and appearance of solutions.
9. Solutions are tested with **test data** to fully exercise all conceivable types of data inputs.
10. Good test data especially checks **boundary conditions**, where the behaviour of a solution should change.
11. The **Gantt chart** you created in Unit 3, Outcome 2 is continued in Unit 4, Outcome 1 to monitor progress and record changes in the original schedule.
12. Gantt charts and project plans should be updated on an ongoing basis to include real events as projects proceed.
13. Each **evaluation criterion** has a corresponding method that answers the question whether the criterion has been satisfied.
14. Evaluate by **objective measurement** whenever possible. Use interviews only to discover people's feelings, opinions and judgements.
15. Evaluate **project plans** to find ways to improve projects in the future.

9780170440875

TEST YOUR KNOWLEDGE

Review quiz

Data visualisations

1 Infographics and data visualisations both use data. How do infographics and data visualisations differ from a written report?
2 How is an infographic different to a dynamic data visualisation?
3 Why does all the data need to be 'cleansed' or wrangled? What happens if it is not?
4 What are some of the features of (**a**) infographics, and (**b**) data visualisations?

Procedures and techniques for managing files

5 What is a 'dashboard' and what does it do?
6 How can a data visualisation be made dynamic?

An effective infographic or data visualisation

7 List three ways in which a graphic solution may be changed to improve accessibility.
8 What is colloquial language, and why should it be avoided in an infographic?
9 What techniques might you use to appeal to younger and to senior audiences?
10 What is 'colour accessibility'? Why is colour an accessibility factor?
11 What types of media can be used in an infographic or data visualisation? What limitations might apply?
12 How can compatibility problems prevent an infographic or data visualisation from conveying its message to a target audience?
13 Explain the main benefits of using cell styles in an Excel worksheet.

Managing data

14 Why is having a second copy on the same hard drive not considered a 'safe' backup?
15 What does the 3-2-1 backup strategy mean?
16 How should a file *not* be named? Explain how a file should be named.
17 Why is 'version control' a problem? How can this problem be overcome.
18 What is a disaster recovery plan? Provide details of your DRP.

Formats and conventions

19 Explain the difference between formats and conventions.

9780170440875

TEST YOUR KNOWLEDGE

Manipulating data

20 List three types of software that could create an infographic or data visualisation for your Data Analytics SAT. Choose one of them and explain why it is the best for your circumstances.

21 Suggest two ways to make a data visualisation more dynamic or interactive.

22 There are three types of validation techniques. Explain the similarities and differences between each of them.

23 What is testing? How is it different to validation? List five types of testing.

24 What is 'good test data'? What other tests might be included for your graphic solution?

Evaluating your solution

25 Distinguish between 'evaluation' and 'testing'.

26 When should evaluation occur?

27 How is evaluation performed? What are 'evaluation criteria', and where do they come from?

Documenting the progress of projects

28 Why should Gantt charts be updated as projects proceed?

29 List three examples each of effectiveness and efficiency evaluation criteria. Give an appropriate evaluation method for each criterion.

30 When should interviews be used during evaluation?

31 Why should project plans be evaluated after a project ends?

SCHOOL-ASSESSED TASK TRACKER

☑ Project plan	☑ Collect complex data sets	☑ Analysis	☑ Folio of alternative designs	☑ Infographic or dynamic data visualisations	☑ Evaluation and assessment	☑ Finalise report or visual plan

9780170440875

APPLY YOUR KNOWLEDGE

Create a data visualisation that presents the information you generated about your research topic and the conclusion this information led you to regarding whether the research question is supported or not. Below are some notes that you should consider when creating this data visualisation.

1. While your data visualisation may not actually be published, it must be suitable for a appropriate audience.
2. Your data visualisation should use appropriate formats for presenting data and information, and use appropriate conventions for each format.
3. You should apply relevant design principles to your data visualisation to enhance its functionality and appearance.
4. Present two or three of your design ideas of the structure and appearance of the data visualisation of your research findings (hand-written notes are satisfactory).
5. Present the details of your data visualisation's design structure and appearance.
6. Describe how you applied manual and electronic validation techniques.
7. Show evidence of testing, such as testing tables, screenshots of functionality testing and test-user feedback.
8. Using a written report or an annotated visual plan, assess how effectively the Gantt chart helped you to manage the development of your research, and explain how the timeline was modified as the project proceeded.
9. Write a self-evaluation of your work during the development of your research, describing how well the finished product satisfied its original aims, and what could be improved if you repeated the project.
10. Describe the procedures you used to manage all your files relating to this SAT.

9780170440875

PREPARING FOR Unit 4 OUTCOME 1

Develop and evaluate an infographic or dynamic data visualisation solution that presents the findings to a research question and assess the effectiveness of the project plan in managing progress.

To achieve this Outcome, you will draw on key knowledge and key skills outlined in Area of Study 1. This Outcome concludes the data acquisition, processing and analysis of the research question formulated in Unit 3, Outcome 2.

Outcome milestones

1. From Unit 3, Outcome 2, use your preferred design solution. This was chosen from two or three feasible alternative design ideas. These other rough sketches or mock-ups of development strategies were without much detail. You can continue to expand on the preferred design idea *and add detail to create a complete design*. Note, however, that these improvements *will not be reassessed*, so do not change the preferred design unless you must.
2. Develop an educational infographic or dynamic data visualisation intended for your target audience that communicates the findings of the research question you developed and researched in Unit 3, Outcome 2.
3. Evaluate the effectiveness of the solution.
4. Assess the effectiveness of your project plan in monitoring your project's progress.

Steps to follow

1. Develop the infographic solution, including:
   - a. building the actual infographic or dynamic data visualisation
   - b. performing a manual and/or electronic validation of the input data
   - c. thoroughly testing that the infographic solution works as intended.
2. Evaluate the effectiveness of your solution using the criteria established during design. Note that this is *not* the same as testing: you should judge how well the finished solution succeeds in satisfying the requirements of the evaluation criteria. This may require interviews with several willing reviewers.
3. In a written report, or an annotated visual plan, assess the effectiveness of your project plan (Gantt chart) in monitoring your project's progress from the start of Unit 3, Outcome 2 to the end of Unit 4, Outcome 1.

Documents required for assessment

1. A solution that communicates the findings of the research question detailed in Unit 3, Outcome 2 as an infographic or dynamic data visualisation.
2. An evaluation of the effectiveness of the solution.
3. An assessment of the effectiveness of the project plan (Gantt chart) in monitoring project progress in a written report or an annotated visual plan.

Assessment

Your teacher will provide you with a more detailed set of assessment criteria before you begin this assessment.

The SAT (comprising Unit 3, Outcome 2 and Unit 4, Outcome 1) will contribute 30% towards your study score.

SCHOOL-ASSESSED TASK TRACKER

☑ Project plan	☑ Collect complex data sets	☑ Analysis	☑ Folio of alternative designs	☑ Infographic or dynamic data visualisations	☑ Evaluation and assessment	☑ Finalise report or visual plan

9780170440875

CHAPTER

6 Information management

KEY KNOWLEDGE

After completing this chapter, you will be able to demonstrate knowledge of:

Digital systems

- characteristics of wired, wireless and mobile networks
- types and causes of accidental, deliberate and events-based threats to the integrity and security of data and information used by organisations
- physical and software security controls for preventing unauthorised access to data and information and for minimising the loss of data accessed by authorised and unauthorised users
- the role of hardware, software and technical protocols in managing, controlling and securing data in information systems
- the advantages and disadvantages of using network attached storage and cloud computing for storing, communicating and disposing of data and information

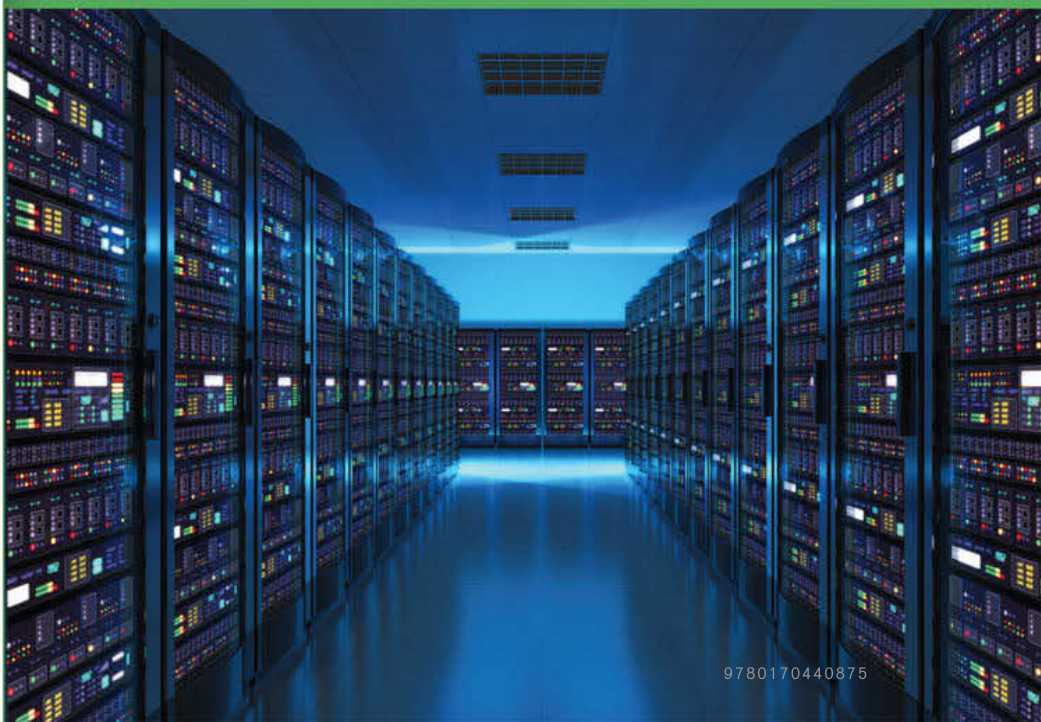
Reproduced from the VCE Applied Computing Study Design (2020–2023) © VCAA; used with permission.

FOR THE STUDENT

Networks are an important part of modern information systems. A network connection is built into nearly every modern computing device. In this chapter, you will learn about the characteristics of networks and the hardware, software and protocols required to control and manage a network. The threats to data and information are also explored along with the security controls that can be used to minimise these threats. Finally, you will learn about the advantages and disadvantages of network attached storage and cloud computing.

FOR THE TEACHER

Networks are an import part of modern information systems and students require an understanding of the characteristics of each type of network, the hardware, software and protocols used to control and manage a network, the threats to data and information in a network environment and the controls that can be used to reduce the chance of data loss due to these threats. Knowledge of both network attached storage and cloud computing is required, including an understanding of the advantages and disadvantages related to both in terms of storage, communication and disposal of data and information. This chapter is based on Unit 4, Area of Study 2, and together with Chapter 7, provides the key knowledge to complete Unit 4, Outcome 2.



Shutterstock.com/Oleksiy Mark

9780170440875

Networks

A **standalone device** is any piece of computing equipment that can perform its function without the need of another device, computer, or connection. The device does not communicate with any other device (Figure 6.1).

A **network** consists of two or more digital system connected devices, using some form of transmission media between them (Figure 6.2).

THINK ABOUT DATA ANALYTICS 6.1

What are some advantages and disadvantages of a standalone computer?



Shutterstock.com/Taner Muhlis Karaguzel

FIGURE 6.1 A standalone device



Shutterstock.com/Dukes

FIGURE 6.2 A network

The purpose of a network is to allow devices to be able to communicate with each other. For successful communication, a network needs:

- a sending device (such as a notebook computer) that initiates a command to transmit data, instructions or information
- a communications device (such as a wireless adaptor inside a notebook computer) to forward packets of data, instructions or information from a sending device via signals carried by a communications channel
- a communications channel or transmission media (such as a cable or radio waves) through which the digital signals travel
- a communications device (such as a wireless router) that receives the signals from the communications channel and forwards the packets to the receiving device
- a receiving device (such as a printer) that accepts the data, instructions or information.

Notebook computers, tablets, smartphones and other sending devices usually have a built-in communications device.

The primary function of a communications device, such as a broadband router, is to transmit data, instructions and information between a sending and a receiving device along a communications channel in digital form.

A digital signal consists of individual electrical pulses that represent the bits grouped together into bytes. Early networks used analogue signals, which consist of a continuous electrical wave. Computers process data as digital signals, so a modem was used to convert between the analogue and digital signals.

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

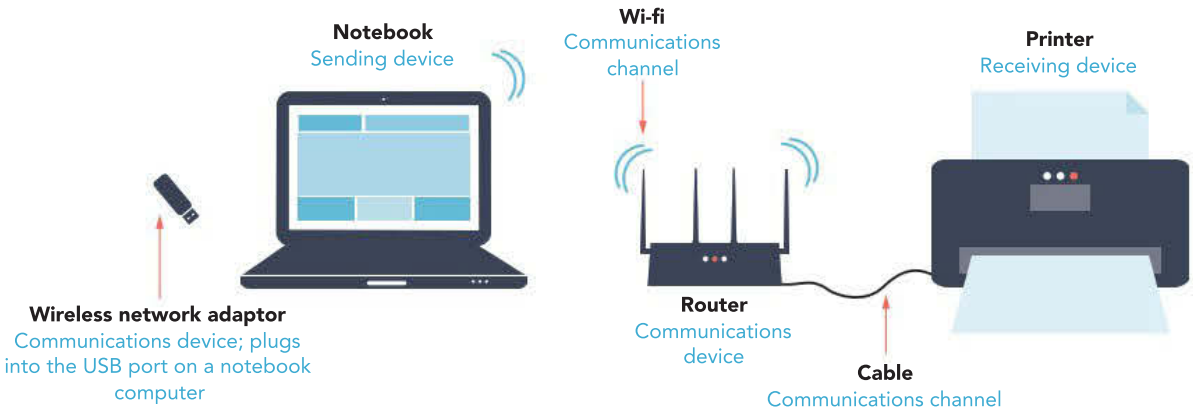


FIGURE 6.3 Sending, communication and receiving devices

Advantages of networks

Advantages of networks include:

- sharing data and information
- allowing communication
- sharing hardware and software.

Sharing data and information

Once devices are connected, they are able to share data and information. Files on one device may be accessed by other devices on a network. Many networks use file servers, where data and information is stored on a central computer, and users of the network can access the files on this central server. The ability to share data and information facilitates collaboration between users and also assists with version control.

Version control is the management of files and software as changes are made. By storing files on a central server, only one version of a file is in existence at a time. As changes are made to the file, the most recent version is stored on the server. As the next user accesses the file, they will be accessing only the most recent version.

Allowing communication

Users on network devices can communicate with each other. This may involve email, video conferencing, online chat or accessing shared calendars. Communication may occur between users within a local area network or with users over a wider network by utilising an internet connection.

Sharing hardware and software

Once devices are connected and can communicate, they are also able to share digital system resources. These resources can include printers, scanners and projectors. For example, each standalone computer would need to connect directly to a printer to be able to produce hard copies of data and information. Once a network is created, only one printer would be required for the network and each device could send requests to the printer.

9780170440875

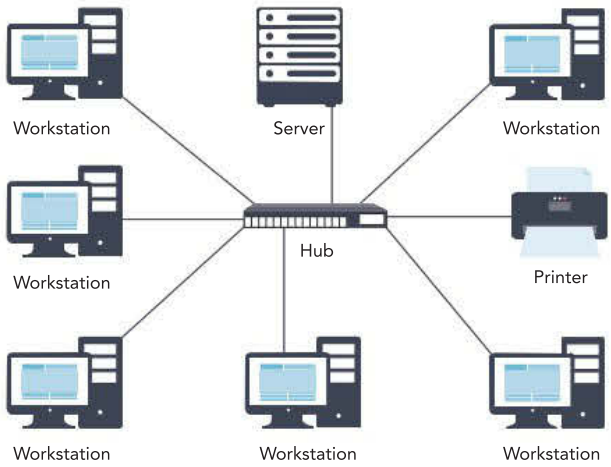


FIGURE 6.4 A network sharing a printer

Each computer or device on the network is considered to be a **node**. Nodes can be connected to a network either via wires or wirelessly.

Types of networks

Networks can often be grouped into two types: a local area network and a wide area network.

Local area network (LAN)

A **local area network (LAN)** is created when two or more devices are connected together in the same geographical area (usually up to a kilometre or two in size). A number of devices connected together in a home, or a group of devices that are connected at a school, are examples of a LAN (Figure 6.5).

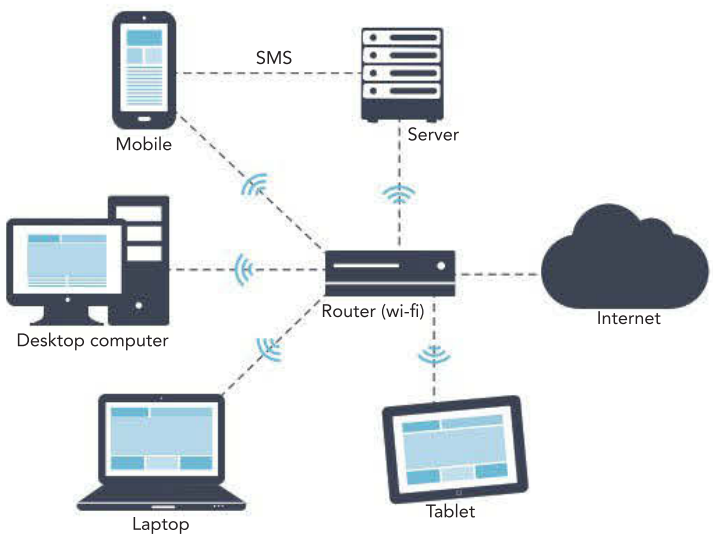


FIGURE 6.5 A local area network with internet connection

The devices are connected to each other without the need of an internet connection, usually using wired transmission media or a wireless connection. A LAN will often have a connection allowing the devices to communicate over the internet.

THINK ABOUT DATA ANALYTICS 6.2

What are other advantages of connecting two or more devices together?

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

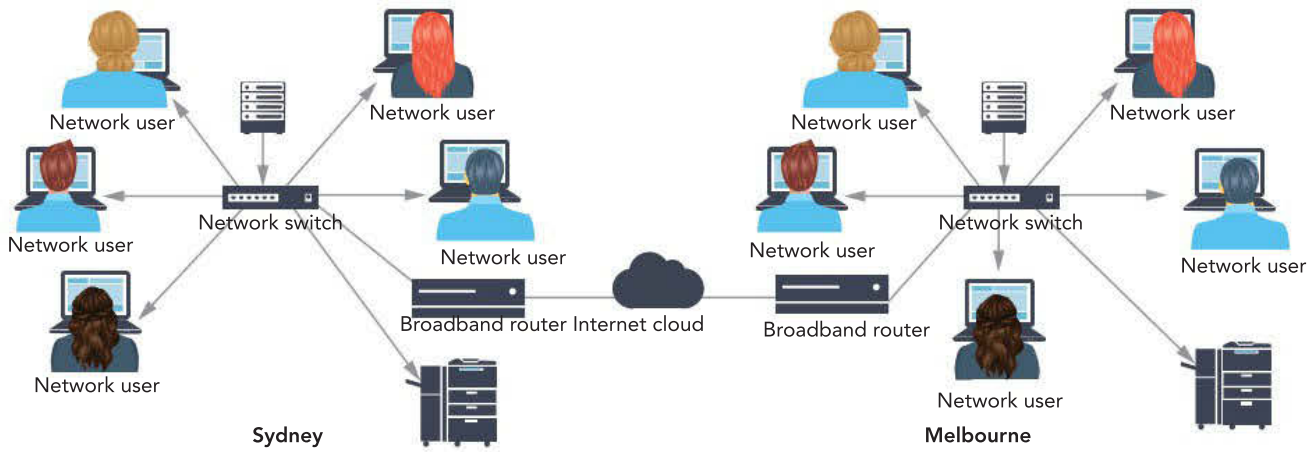


FIGURE 6.6 A wide area network using an internet connection

THINK ABOUT DATA ANALYTICS 6.3

Explain why an organisation that has offices in two different cities would use third-party transmission media rather than install their own dedicated connection.

Wide area network (WAN)

A **wide area network (WAN)** network is located in more than one geographical location. Often two LANs that are based in different locations are able to communicate with each other using an internet connection. For example, two offices, each with its own LAN, can communicate with each other using an internet connection.

One difference between a LAN and a WAN is that a WAN generally uses third-party transmission media to connect the two locations, often an internet connection. Third-party transmission media refers to media that belongs to another organisation. This will often be a telecommunications organisation, such as Telstra, Optus or Vodafone. Once an organisation uses third-party transmission, it starts to lose some control over the network and the potential of data threats increases.

Network architecture

Network architecture refers to the layout of a network, including the hardware, software, protocols, and transmission media it uses. Two common types of network architecture are peer-to-peer networks and a client/server networks.

Peer-to-peer network

In **peer-to-peer networks**, all the devices are equal. All devices store and share data with other devices. Most home networks are set up as peer-to-peer networks.

Client/server network

A **client/server network** has a central **server**. A server is a large computer that stores data and software. The client devices request data stored on the server. The server also assists in controlling which users can access which data and services, and has security measures to protect data and information.

Most schools have a client/server network installed. This allows staff and students to store data on shared drives. Another advantage of a client/server network is that servers, which are responsible for managing particular tasks, free up resources on client devices, allowing performance to be improved across the network.

Types of servers include the following.

- A file server stores files (data and information) that client devices can access.
- An email server stores the network users' sent and received emails.

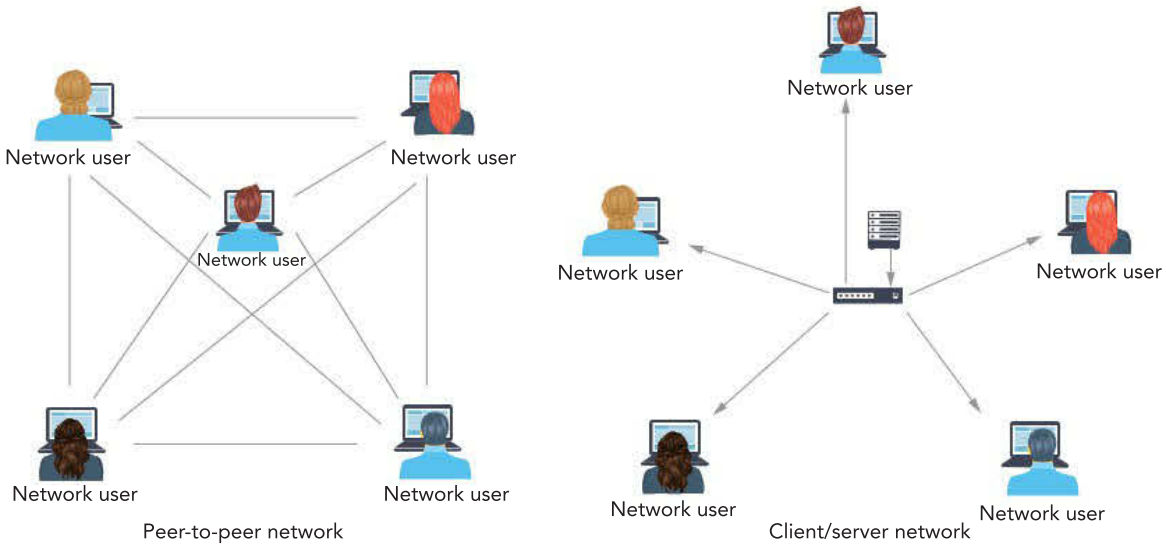


FIGURE 6.7 A peer-to-peer network and client/server network

- A proxy server sits between a LAN and the internet. Requests from all network users to access websites will be sent to the proxy server, and the proxy server will send the request across the internet.
- A web server stores files related to a website. When a user requests a web page by typing in a URL, they are directed to the web server where that page is located.
- A database server stores the data related to a database management system.
- A print server manages requests for printing on a network, creates print queues and sends print jobs to the relevant printer.

Dedicated and virtual servers

Traditionally, servers were physical devices consisting of large amounts of storage space known as dedicated servers. Each device was confirmed to perform one particular server task (such as a file, email or print server).

In recent years, the use of virtual servers has increased. Virtualisation allows one physical dedicated server to be divided into multiple virtual servers using virtual server software to take full advantage of the processing power of the physical server. Each virtual server runs independently of the other virtual servers stored on the physical component and acts as a unique physical device.

This virtualisation reduces the number of physical servers required by an organisation and streamlines the network infrastructure. This allows applications and servers to be deployed more quickly and services to be more available to users of a network. Virtual servers are also less expensive to set up and easier to manage than a network with a number of physical dedicated servers. They are also more efficient to run in terms of the power consumed.



iStock.com/Grassetto

FIGURE 6.8 A server

Wired networks

In a **wired network**, the devices are connected using a physical cable. Two types of cabling used are fibre-optic and copper cabling. Fibre-optic is often used as the backbone of a network, while copper cabling is used to connect individual nodes together.

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

Advantages of wired networks compared to wireless networks include:

- faster data transfer speeds
- better security (in regards to data interference or theft)
- more reliable connection.

Disadvantages of wired networks compared to wireless networks include:

- lack of mobility
- installation
- maintenance.

Faster data transfer speeds

Although wireless and mobile network data transmission rates are continuously improving, the speeds generated by wired networks are generally viewed as superior.

Better security

Wired networks are considered more secure than wireless networks. In a wired network, it is much more difficult for a hacker to intercept the data (particularly in a local area network). There is also less chance that data is lost due to signal interference.

More reliable connection

Not only are wired networks viewed as more reliable in terms of data interference, but but they're also considered to have more consistent data transfer when compared to wireless networks. In wireless networks, the strength of the signal may vary due to a range of factors, which may in turn affect the data transfer speeds.

Lack of mobility

Since wired networks generally have fixed data points, which are used to connect devices, this reduces the flexibility to where devices must be located to stay connected to the network. Given this makes devices stationary, this limits the ability of the user to roam while continuing to use their device.

Installation

The cost of installing a wired network is considerably more expensive than installing a wireless network. The cost of purchasing cabling and having the cabling installed, as well as the time and effort to compete the installation, is much greater than installing a wireless access point.

Maintenance

From time to time wired network cabling may stop working correctly or the organisation requires new data points or rearranging of existing data points. Just as the cost of installation can be significant, the cost of maintaining and making changes to the existing wired network is also time-consuming and costly.

Wired transmission media

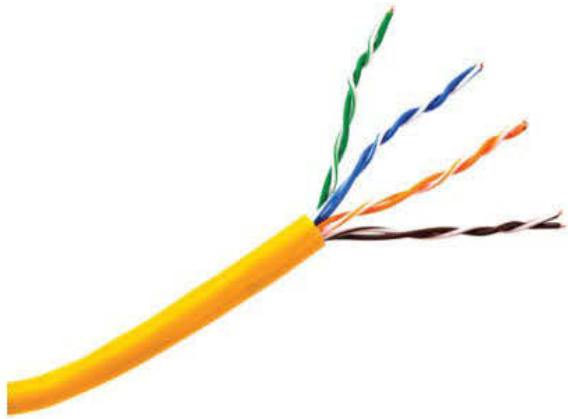
Two common types of wired transmission media are unshielded twisted pair cabling and fibre-optic cabling.

9780170440875

Unshielded twisted pair

Unshielded twisted pair (UTP) cabling is used in most wired networks. Wires are twisted into pairs to reduce interference. Unshielded twisted pair is also referred to as CAT (such as CAT 5) or Ethernet cable. UTP cabling is generally used for distances of up to 100 metres. The data transfer rate of UTP has increased over time as each new category of cable is released.

One weakness of UTP cabling is that the signals travelling through the cable can be affected by interference, which can corrupt the data. UTP cabling is normally found within a local area network to connect network devices that are situated relatively close to each other.



Shutterstock.com/Zakhar Mar

FIGURE 6.9 An unshielded twisted pair (UTP) cable

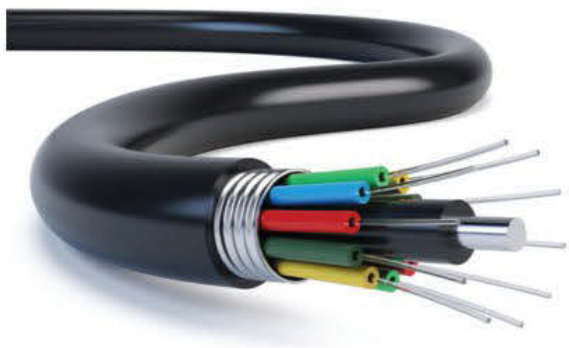
TABLE 6.1 Categories of UTP cabling

Category	Data rate	Distance
CAT 5	100 Mbps	100 metres
CAT 6	1 Gbps	100 metres
CAT 7	10 Gbps	100 metres
CAT 8	40 Gbps	100 metres

Fibre-optic cabling

Fibre-optic cabling is a wired transmission media that contains shards of glass that reflect pulses of light generated by small lasers or light-emitting diodes (LEDs).

Because glass is used to reflect light (which contains the data), fibre-optic cabling has a much higher data transfer rate than UTP. While UTP has a maximum reach of around 100 metres, fibre-optic cabling can be used to transmit data over long distances (up to approximately 2 kilometres). Fibre-optic cabling is immune to interference, meaning that data cannot be corrupted as it travels between devices.



Shutterstock.com/Sashkin

FIGURE 6.10 Fibre-optic cabling

Fibre-optic cabling is used as the backbone of the National Broadband Network. Most premises are connected to the NBN using a fibre to curb connection, where fibre-optic is connected to the front of the house, while unshielded twisted pair is used from the premises to the curb.

Fibre-optic cabling is very delicate and can easily be damaged. One limitation of fibre-optic cabling is that its data transfer speeds are reduced when the cabling is bent. Bending the cable increases the potential for the glass within the cable to become damaged, and therefore, fibre-optic cabling is normally used between buildings and often installed underground to prevent the need to bend the cabling and avoid the chance of the cabling becoming damaged. Fibre-optic cabling is also more expensive than UTP cabling.

6.4 THINK ABOUT DATA ANALYTICS

Why do you think fibre-optic cabling is not normally used within a single room in a building?

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

Wireless networks

A **wireless network** is a computer network that uses wireless data transfer between network devices. Wireless networks generally use radio communication to transmit data. They include:

- wireless local area networks (WLANs)
- mobile networks
- satellite networks
- microwave networks
- infrared
- Bluetooth.

A popular type of wireless network is a wireless local area network, commonly known as as wi-fi. WLANs use the **802.11 protocol** wireless standard to create a wireless connection between devices on a network.

WLANs use a **wireless access point (WAP)** or **router** to broadcast a signal using WAP or WEP encryption to send and receive signals between devices on the network.

TABLE 6.2 Categories of 802.11 wireless standards

Category	Data rate	Range (indoors)
a	54 Mbps	35 metres
b	11 Mbps	35 metres
g	54 Mbps	40 metres
n	600 Mbps	70 metres
ac	1 Gbps+	35 metres

THINK ABOUT DATA ANALYTICS 6.5

Think of areas where mobile phone towers do not overlap. What is the consequence to users if they are in an area where they are not in reach of a mobile phone tower?

A **mobile network** (also known as a cellular network) uses telecommunication networks to allow users to communicate using their mobile device. A mobile network consists of a number of mobile phone towers (or base stations) that send and receive signals. Mobile phone tower data coverage often overlaps, allowing users to stay connected to the network as they roam an area.

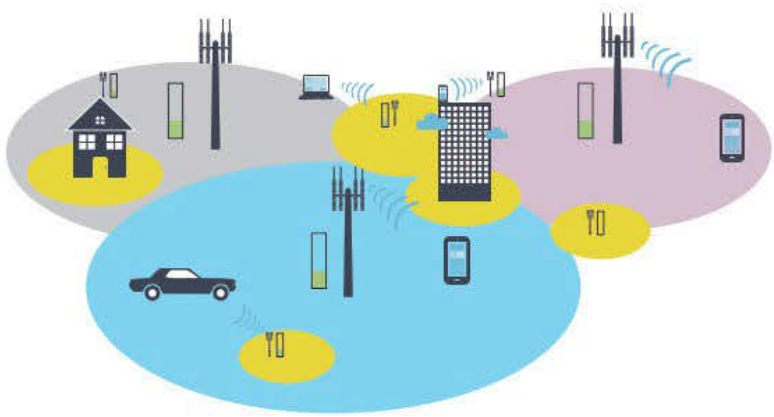


FIGURE 6.11 A mobile network showing cell tower coverage

Infrared transmission uses the same technology as TV and video remote controls. It is usually effective over short distances (up to about five metres), although the data transfer rate is slow compared to using cables. Many hand-held computers have infrared ports that can

9780170440875

communicate easily with printers or laptops. Thus, material can be backed-up quite easily and updated from another computer. Infrared transmission uses light waves and requires line-of-sight access.

Bluetooth uses short-range radio waves to transmit data among Bluetooth-enabled devices. These devices contain a small chip that allows them to communicate with other Bluetooth-enabled devices. Examples of these devices include desktop personal computers, notebook computers, hand-held computers, mobile telephones, fax machines and printers. To communicate with one another, they must be within a specified range (about 10 metres, but the range can be extended to 100 metres with additional equipment). A popular use of Bluetooth is to enable hands-free communication on mobile phones. Most cars are now sold with a built-in Bluetooth station that the user can synchronise with their mobile phone. Bluetooth and wi-fi communications technologies both use radio signals.

The range of each mobile phone tower can vary depending on its location and the number of other towers located in the area. In large cities, each tower may only have a short range (for example, 2 or 3 kilometres), but often these towers overlap with numerous other towers. In more isolated areas the tower may have a larger coverage area (for example, approximately 10 to 50 kilometres depending on the terrain).

A user's mobile device will connect to the nearest base station. Each base station is then connected to a digital exchange where the communication is sent over a wired network.

Mobile networks allow the user to send and receive voice, data, images, and text messages. Because most (if not all) mobile networks belong to a telecommunication organisation, there is usually a cost associated with transmitting data using these networks.

TABLE 6.3 Generations of cellular transmission

Generation	Year	Data rate
2G	1993	14.4 Kbps
3G	2001	3.1 Mbps
4G	2009	100 Mbps
5G	2020*	20 Gbps

*approximately

The Internet of Things

The **Internet of Things** is a network formed by 'smart devices' such as mobile phones, wearable devices, headphones, electricity meters, household whitegoods, connected cars, home assistants and other devices that have an 'on-off' switch to the internet and/or to other Internet of Things devices.

These devices must have a method of connecting to the internet, whether by wireless or wired technology.



Shutterstock.com/elenabsl

FIGURE 6.12 The Internet of Things

THINK ABOUT DATA ANALYTICS 6.6

Create a list of devices commonly found within a household that can be connected to the internet.

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

A White Paper published by IBM listed the following causes of data loss.

- User error 32%
- Hardware malfunction 30%
- Software malfunction 14%
- Malware 7%
- Natural disasters 3%

Threats to data and information

A **threat** refers to anything that has the potential to cause harm to data and information stored and communicated between information systems. A threat is something that may or may not happen, but has the potential to result in the loss, theft or damage of data and information. Threats can be classified into three groups: accidental, deliberate and event-based threats.

Accidental threats

A threat is classified as an **accidental threat** if the potential to cause harm to data and information is unintentional and unexpected (meaning, it is an accident).

User error

User error is any error that has been caused by the user, rather than due to the hardware or software not working as expected. A user deleting files that contain data and information they still require is an example of user error.

User errors occur due to individuals lacking knowledge in the use of an information system or from users not paying enough attention when using an information system. It can also be caused by a failure to follow procedures, such as running a daily backup or using correct file-naming conventions.

THINK ABOUT DATA ANALYTICS 6.7

Think about other examples of user error besides inadvertently deleting files.

Power loss

Power loss (not to be confused with a power surge) can occur any time the power for an information system is lost. This could be due to an electricity outage used to provide the supply power to a system, a power cable becoming loose (or displaced by accident) or the battery of a device running out of power.

Hardware or software malfunction

Hardware or software malfunction occurs if the hardware or software stops functioning for some reason. Data loss can occur if the hardware or software of an information system or device stops working correctly. This can involve the electronic circuits within a hardware device ceasing to operate as expected, or disks within a hard disk drive becoming damaged, leading to the data no longer being able to be retrieved.

Different levels of hardware malfunction cause various levels of data loss. For example, the failure of a graphics card or RAM may result in the loss of open data when the device resets, while the failure of a hard drive may also mean the loss of saved files. Other examples of hardware malfunction include a central processor unit overheating, a blank monitor, a jumpy mouse, port connection issues or an unresponsive keyboard.

Software malfunction occurs when bugs or glitches start to appear in software that previously seemed to be working as expected. This could be due to incomplete testing before the software release, an error in the code caused by a recent upgrade, or new circumstances that the software could not handle. An improper shutdown may lead to the software becoming corrupt and no longer being able to process data.

Hardware loss

Hardware loss occurs when a device or hardware component is accidently misplaced. This could simply be a user leaving their smartphone on a train or a hardware device being thrown away without realising that it contained data that was still required.

9780170440875

Deliberate threats

A threat is classified as deliberate if it was created on purpose to cause loss or damage to data or an information system. These differ from accidental threats, which are the result of a misfortune, and event-based threats that occur as a result of nature. A wide range of **deliberate threats** fall into the broad category of malware.

Malware

Malware, short for malicious software, is designed to either damage, disrupt or gain unauthorised access to an information system. The term malware is used to group together a range of software threats to data and information. Table 6.4 describes the types of malware.

TABLE 6.4 Types of malware

Threat	Description
Adware	Software designed to automatically deliver advertisements to a user's computer. Adware is often bundled together with free trial versions of software that a user may download. Adware is sometimes bundled together with spyware.
Bot	A bot is a software application that performs a specific task autonomously. A bot is often created to perform a repetitive task (for example, web crawler fetches files from a web server). Bots can also be created for malicious purposes, including for DDoS attacks, to create adware or distributing other malware such as Trojans and viruses.
Bug	In terms of software, a bug is an error in the source code that prevents software producing the expected outcome. In terms of malware, software can be written to alter the source code of a program to ensure that the software does not function as expected.
Keylogger	Software that, once installed, records all of the keystrokes made on a device's keyboard. This data can then be sent to a remote location and analysed to identify files and websites visited and usernames and passwords used.
Ransomware	Ransomware uses encryption to encode all of the data on a device or network without the user's knowledge. To decrypt the data, the users are asked to pay money (a ransom) to unknown people who were responsible for the encryption in the first place. Often ransomware is downloaded in conjunction with another file or through a user clicking on a link in an email.
Rootkit	A rootkit is software that is designed to control a device from a remote location, without the user's knowledge. Once installed, it is possible to execute files, edit or steal data and information, or use the device as a bot.
Spyware	Spyware is software that collects data about a user's activity and sends that data to another location, without the user's knowledge. Spyware can also include functions that allow network or security settings to be changed on a user's device. Similar to other malware, spyware can be downloaded inside a Trojan or after clicking on a link in an email.
Trojan	A Trojan is software that appears to be 'normal' but contains hidden malware. Once the user installs the 'normal' software, the malware hidden within is also (secretly) installed on the device.
Virus	A virus is software designed to cause some type of negative effect on a device or network. When a virus is executed, it normally duplicates itself in a range of locations on a device. A virus may be designed to damage, steal, modify or corrupt data.
Worm	Once installed, a worm is software that self-replicates. The purpose of the worm is typically to replicate itself over and over again, causing the performance of a device or network to slow down as the software starts to drain all of the system resources. Some worms also contain a 'payload' that may cause some type of damage to either hardware or software.

THINK ABOUT DATA ANALYTICS 6.8

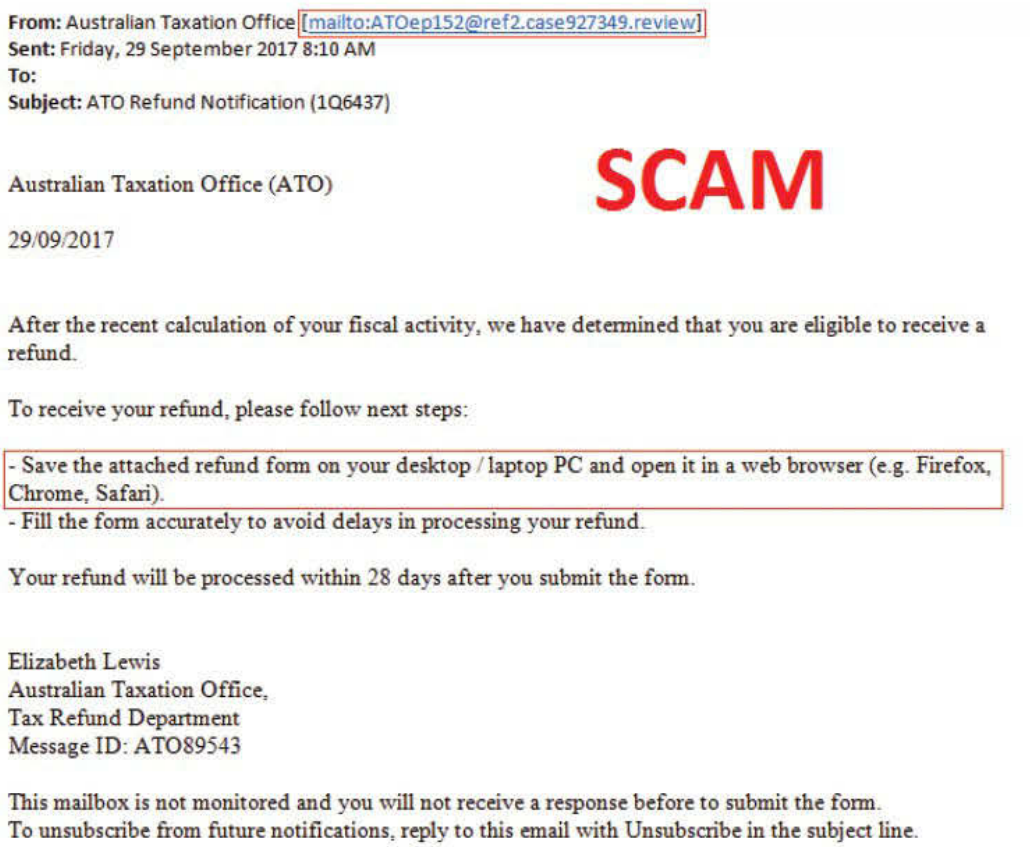
Think of some other types of software that may be classified as malware.

9780170440875

Phishing

Phishing involves the sending of hoax emails or messages to users pretending that the communication is from a legitimate source, such as a bank, the ATO or an online shop.

The purpose of the hoax email is to trick the user into revealing sensitive information (for example, username and password, bank account number, date of birth, and so on) thinking that they are providing the information to the legitimate source.



From: Australian Taxation Office [mailto:ATOep152@ref2.case927349.review]
Sent: Friday, 29 September 2017 8:10 AM
To:
Subject: ATO Refund Notification (1Q6437)

Australian Taxation Office (ATO)

SCAM

29/09/2017

After the recent calculation of your fiscal activity, we have determined that you are eligible to receive a refund.

To receive your refund, please follow next steps:

- Save the attached refund form on your desktop / laptop PC and open it in a web browser (e.g. Firefox, Chrome, Safari).
- Fill the form accurately to avoid delays in processing your refund.

Your refund will be processed within 28 days after you submit the form.

Elizabeth Lewis
Australian Taxation Office,
Tax Refund Department
Message ID: ATO89543

This mailbox is not monitored and you will not receive a response before to submit the form.
To unsubscribe from future notifications, reply to this email with Unsubscribe in the subject line.

FIGURE 6.13 An example of a phishing email

© Australian Taxation Office for the Commonwealth of Australia

THINK ABOUT DATA ANALYTICS 6.9

List some signs that help to identify an email as a phishing attempt.

Once the person responsible for the fake email receives sensitive information, they attempt to use that information to commit some type of fraud, such as withdrawing money from the user's bank account, applying for a loan or buying goods online.

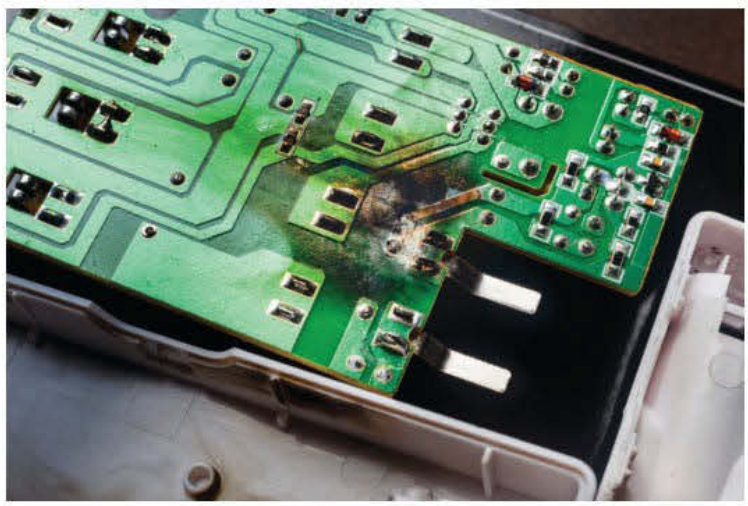
Event-based threats

An **event-based threat** is a threat to data and information that is a result of a natural event. This includes fires, floods, heatwaves, storms and earthquakes. As a result of the natural event, damage may occur to the hardware and software of a device or information system.

Power surge

A **power surge** is considered one type of event-based threat. A power surge is an abnormally high voltage of electricity received at a location in a short amount of time. Power surges are usually caused by a lightning strike, which produces the abnormal high voltage. These high voltages can cause an arc of electrical current within the device. The heat generated

in the arc causes damage to the electronic circuit boards and other electrical components (Figure 6.14). Repeated smaller power surges may also slowly damage circuit boards.



Shutterstock.com/Arnain

FIGURE 6.14 A damaged circuit board due to a power surge

Physical and software security controls

A **security control** is a measure designed to protect data and information from threats, either accidental, deliberate or event-based. Security controls may consist of hardware, software or procedures that are used to assist in the protection of an information system. Security controls by themselves, however, do not guarantee that all data and information will be fully protected. Rather, they reduce the chance of unauthorised access and/or data loss. Security controls can be classified into two groups: software controls and hardware controls.

Software security controls

Software security controls are software-based measures that are used to protect data and information.

Usernames and passwords

A combination of a **username and password** is one of the most widely used security controls. Each user of an information system is allocated their own unique username while, in most cases, a user is able to create their own password. Collectively, a username and password is often referred to as a login.

Some information systems require users to create a password that meets particular requirements. For example, the password:

- must consist of at least eight characters
- must contain at least one upper case letter
- must contain at least one lower case letter

9780170440875

- must contain at least one symbol
- must not contain spaces
- cannot be 'password'
- cannot be the same as the username.

Some information systems limit the number of attempts a user has to login successfully. If the user inputs the incorrect username or password a number of times, they may be locked out of the system for a specific time period or until they contact the administrator and reset their credentials. Other information systems allow a user unlimited attempts to gain access. These systems are particularly vulnerable to brute force attacks.

Brute force attacks

A **brute force attack** involves making repetitive attempts at accessing an information system using a variation of possible usernames and/or passwords, although often an attempt will be based on an established username such as an email address.

A bot can be developed and deployed to guess the password that matches a given username. On the first attempt, the password input could be 'a', then 'b', then 'c', right through to 'z', then the next attempt would be 'aa', then 'ab', and so on.

The table below shows the number of attempts required (on average) to guess a password of each particular length (based on the ASCII character set having 128 characters).

A bot is software that performs a specific task autonomously.

Password length	Required attempts (on average)
1	64
2	8192
4	134 217 728
8	36 028 797 018 963 968

THINK ABOUT DATA ANALYTICS 6.10

Why is two-factor authentication now commonly used with username and password?

Because it is possible to create a bot that can make hundreds of thousands of attempts per second, even passwords that are eight characters long can easily be accessed by hackers. A number of organisations now require their users to create a passphrase as the password. A passphrase consists of a number of random words that are joined together. An example is shown here:

boat1yesterdaymonsterpurple#

This creates a password length that would take too long for hackers to continue attempting a brute force attack, while at the same time making it possible for the user to remember the password. Including numbers and symbols in the passphrase decreases the chance of a brute force attack since it removes the threat of a 'dictionary attack'.

Access logs

Access logs can be used on a network, information system and web server. An access log is simply a list of activities performed.

In terms of a network, it will keep a record of what files were accessed, by whom and the time of access. In an information system, an access log will keep a record of each user who logged into the system and the functions they performed. In a web server, an access log will keep a list of requests for files of a particular website. The IP address, date, time, filename, browser and operating system used to access each file will be recorded.

9780170440875

An access log alone will not protect data and information from threats. Rather, it will provide data about the actions of users and assist in identifying possible areas of concern.


User Access Log

Simple Filter Advanced Filter Find

Filter By: Date Matches: *

Showing 1 - 10 of 21,865 records

#	Date	Time	User Name	Origin Host	Action	Object	Severity	Activity	Message
1	2011/11/18	06:15:46	backendadm		Logon	User	INFO	SecurityActivity	Login successfully.
2	2011/11/18	06:15:46	backendadm		Logon	User	INFO	SecurityActivity	Login successfully.
3	2011/11/18	06:15:46	backendadm		Logon	User	INFO	SecurityActivity	Login successfully.
4	2011/11/18	06:12:09	admin	10.65.201.96	Logon	User	INFO	SecurityActivity	Login successfully.
5	2011/11/18	06:12:04	admin	10.65.201.96	Logoff	User	INFO	SecurityActivity	Logoff.
6	2011/11/18	06:11:35	backendadm		Logon	User	INFO	SecurityActivity	Login successfully.
7	2011/11/18	06:11:35	backendadm		Logon	User	INFO	SecurityActivity	Login successfully.
8	2011/11/18	06:11:35	backendadm		Logon	User	INFO	SecurityActivity	Login successfully.
9	2011/11/18	06:10:34	backendadm		Logon	User	INFO	SecurityActivity	Login successfully.
10	2011/11/18	06:10:34	backendadm		Logon	User	INFO	SecurityActivity	Login successfully.

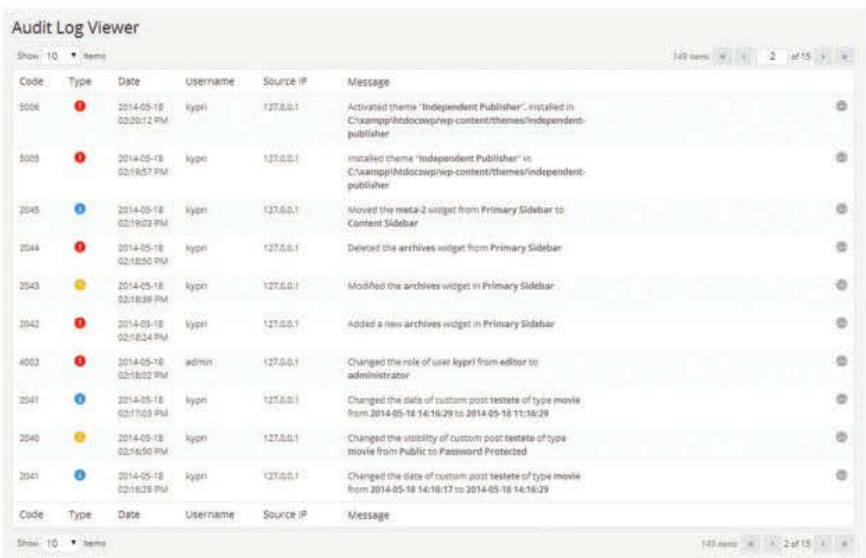
Rows per page: 10 Page 1 of 2187

Courtesy of Cisco Systems, Inc. Unauthorized use not permitted. https://www.cisco.com/c/en/us/td/docs/net_mgmt/prime/fulfillment/6-1/user/guide/prime_fulfill/admin_useraccesslog.html (accessed July 2019)

FIGURE 6.15 An access log

Audit trails

While an access log is used to record individual actions that occur (for example, a user logs in, a file is accessed, a web page is viewed, and so on) an **audit trail** involves establishing a sequence of actions. This can involve piecing together data collected from a number of access logs and may also include data from other sources. Whereas an access log provides static data about system usage, an audit trail provides a dynamic view because it shows connections between actions that are contained within the logs.


Audit Log Viewer

Code	Type	Date	Username	Source IP	Message
5006		2014-05-18 03:20:12 PM	kypri	127.0.0.1	Activated theme 'Independent Publisher', installed in C:\xampp\htdocswp\wp-content\themes\independent-publisher
5005		2014-05-18 02:19:57 PM	kypri	127.0.0.1	Installed theme 'Independent Publisher' in C:\xampp\htdocswp\wp-content\themes\independent-publisher
2045		2014-05-18 02:19:03 PM	kypri	127.0.0.1	Moved the meta-2 widget from Primary Sidebar to Content Sidebar
2044		2014-05-18 02:18:50 PM	kypri	127.0.0.1	Deleted the archives widget from Primary Sidebar
2043		2014-05-18 02:18:39 PM	kypri	127.0.0.1	Modified the archives widget in Primary Sidebar
2042		2014-05-18 02:18:24 PM	kypri	127.0.0.1	Added a new archives widget in Primary Sidebar
4003		2014-05-18 02:18:02 PM	admin	127.0.0.1	Changed the role of user kypri from editor to administrator
2041		2014-05-18 02:17:03 PM	kypri	127.0.0.1	Changed the date of custom post testete of type movie from 2014-05-18 14:16:29 to 2014-05-18 11:16:29
2040		2014-05-18 02:16:50 PM	kypri	127.0.0.1	Changed the visibility of custom post testete of type movie from Public to Password Protected
2041		2014-05-18 02:16:29 PM	kypri	127.0.0.1	Changed the date of custom post testete of type movie from 2014-05-18 14:16:17 to 2014-05-18 14:16:29

WP Security Audit Log

FIGURE 6.16 An audit trail

An audit trail may include data about when a user logged on, files and resources accessed, if any changes were made, and if any files were deleted or sent for print. Similar to an access log, an audit trail on its own does not stop threats to data and information. It provides details about activities on a network and allows issues to be investigated.

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

Access restrictions

Access restrictions (also referred to as permissions) are used to limit the users or groups of users who can access particular files or use particular functions of an information system. For example, only 10 users may be granted access to a particular file. Of those, only five are allowed to edit the file and only one user is given access to delete the file. Access restrictions reduce the chance of unauthorised users accessing the data and information and reduce the chance of data loss because only limited users can delete files.

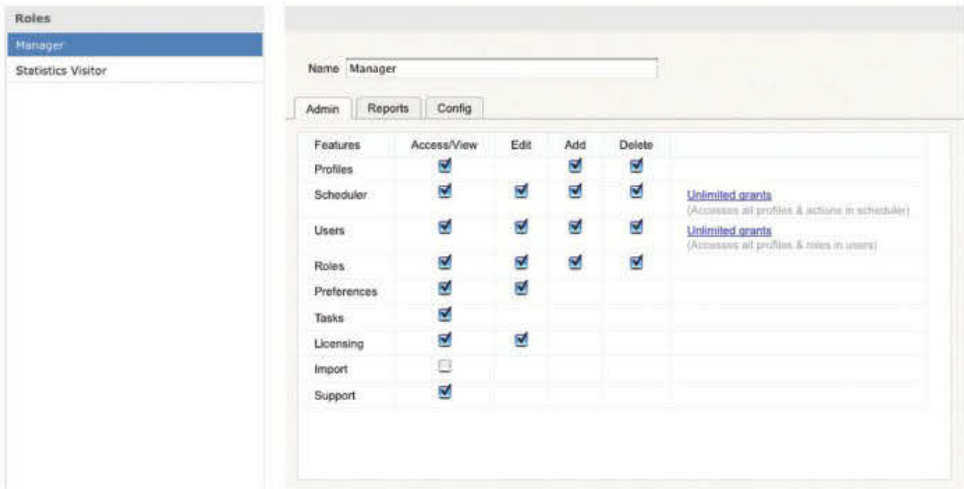


FIGURE 6.17 Access restrictions

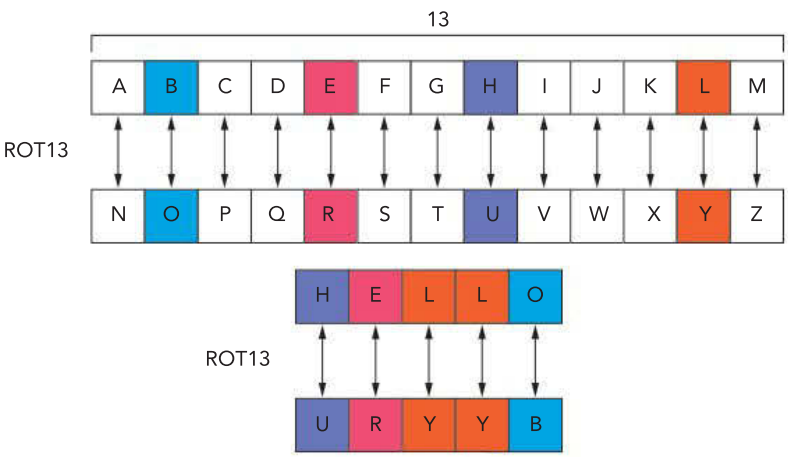


FIGURE 6.18 A simple encryption method

Encryption

Encryption is a security control that is used when communicating data. It can also be used to protect stored data. Encryption does not stop unauthorised users gaining access to data and information; instead, it involves encoding or changing data so if the data was read by an unauthorised user they would not be able to understand it.

A credit card number used to purchase goods online is an example of data that would benefit from encryption.

Figure 6.18 shows a basic method used to encrypt data. The alphabet is broken into two halves. Each letter is paired together with another letter. When data is encrypted, each letter in the data is replaced by its pair.

In the example given, the data 'HELLO', once encrypted, would be transmitted or saved as 'URYYB'.

9780170440875

Private and public keys

Two types of encryption are symmetric and asymmetric encryption.

Symmetric encryption (Figure 6.19) requires both the sender and the receiver to use the same **private key**. This key is only known by a user that has the key installed on their device.

The key is used to encrypt the data packet by the sender and the same key is needed by the receiver to decrypt that data packet. A private key is also referred to as a secret key.

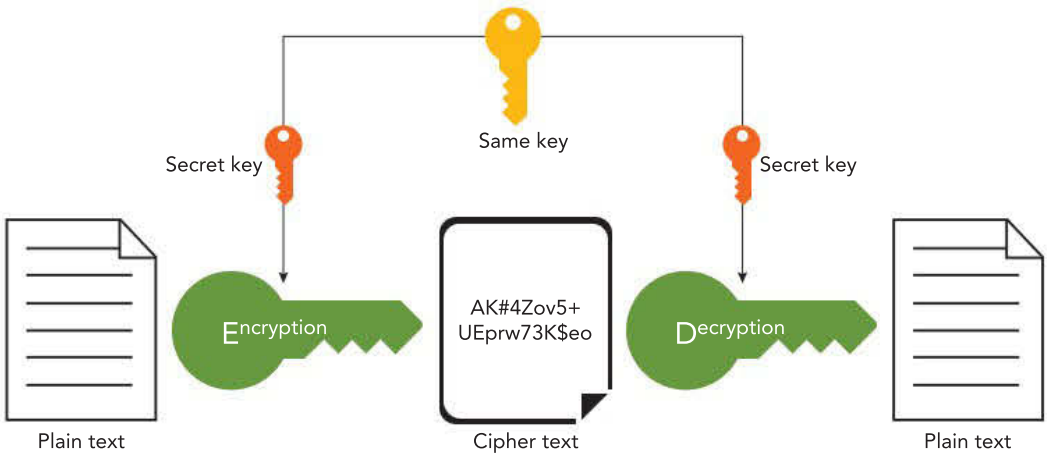


FIGURE 6.19 Symmetric encryption

Asymmetric encryption (Figure 6.20) requires the user to install two keys on their device, a private and a public key. A **public key** is made available to other devices and is used to encrypt data packets that are being sent to the user, while the private key is used on the receiving device to decrypt the data.

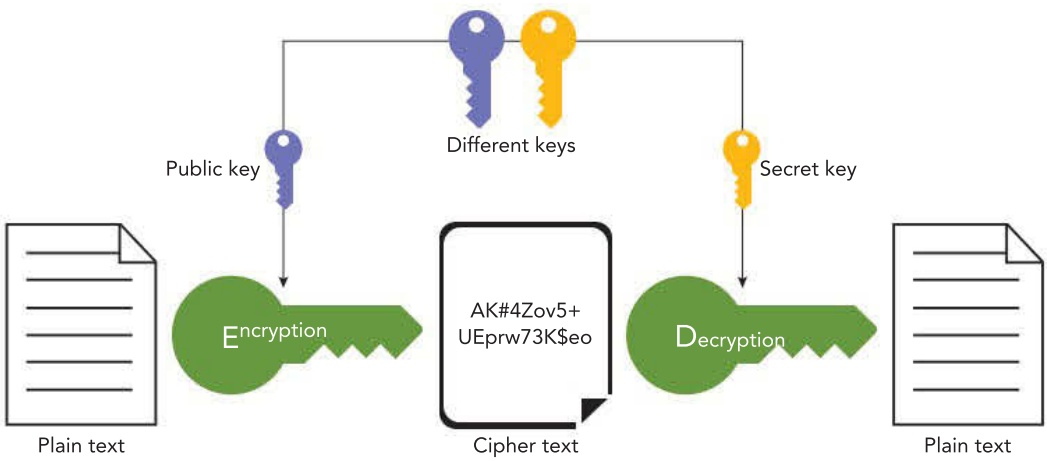


FIGURE 6.20 Asymmetric encryption

The algorithm used by the two keys are related, so that data encrypted by a public key can only be decrypted by a private key.

THINK ABOUT DATA ANALYTICS 6.11

How would the data 'YESTERDAY' be stored after it was encrypted using the method shown in Figure 6.18? How would the data 'CNFFJBEQ' appear after it was decrypted?

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

Firewalls

A **firewall** is software that monitors all incoming (and outgoing) internet traffic to and from a local area network. The firewall examines the contents of each data packet and determines if the data packet should be allowed access to the local area network.

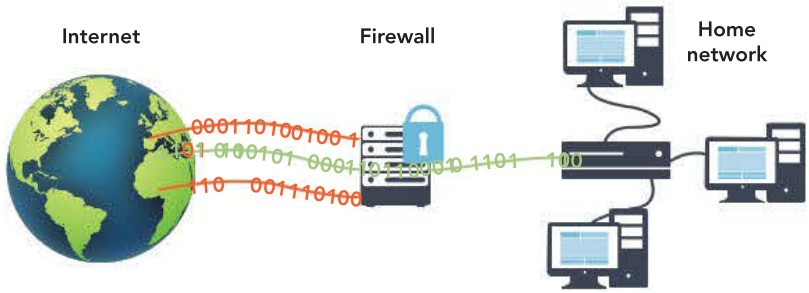


FIGURE 6.21 A local area network with a firewall on an internet connection

A firewall scans the internet connection searching for data packets that have not been requested by the network. These unrequested packets may be hackers trying to gain unauthorised access to the network or internal network users trying to access forbidden material from external locations. Unauthorised data packets are blocked, while authorised packets are allowed to pass through the firewall (Figure 6.21).

System protection software

System protection software contains a number of smaller software applications, that combined, will assist in protecting an information system or device from a range of threats, in particular, malware. System protection software scans a network or device, searching for malicious software. If malware is detected, it is either deleted or quarantined.

System protection software includes:

- firewall software
- content filtering
- backup functions
- cloud storage.

There is a wide range of system protection programs available. Some software is available to download for free, others require a licence to be purchased, while others require a periodic subscription.

The developers of system protection software constantly release software updates that contain revised virus definitions. This ensures that the software is continually providing protection against newly discovered threats.

THINK ABOUT DATA ANALYTICS 6.12

What other features or functions might be found in system protection software?

Security protocols (TLS and SSL)

A **protocol** is a set of rules or guidelines outlining how data will be communicated between devices on a network. A **security protocol** is a set of rules or guidelines for communicating data securely that uses encryption to encode the data so if it is intercepted by an unauthorised user that user would not be able to understand the content of the data.

9780170440875

The web security protocol HTTPS employs encryption to protect data that is communicated. HTTPS creates a secure connection between a web browser and a server over the internet. The data is encrypted to stop 'man in the middle' attacks. HTTPS is based on HTML, using both SSL and TLS protocols.

Secure Sockets Layer (SSL) uses certificates to create a secure connection between a server and a device. The certificate is installed on the web server, and a copy of the certificate is sent to a device. The certificate is used to encrypt and decrypt the data. Transport Layer Security (TLS) is a later version of SSL. It creates a secure connection between a device and a server using certificates.

Hardware security controls

Hardware security controls are hardware based measures that are used to protect data and information. These measures typically consist of zoned security strategies, barrier techniques, biometrics and backing up data.

Zoned security strategies

Zoned security strategies involve breaking down a network into discrete sections or zones.

The purpose of zones is that damage caused by security threats can be limited to one particular area or zone. If a device in one zone is affected by malware, the entire zone can be quarantined, allowing other zones on the network to operate unaffected.

Each zone will use access rights to grant access to a particular user. It will use firewalls to help restrict the data packets entering and leaving the zone and may even use biometrics to restrict physical access to a zone to certain users.

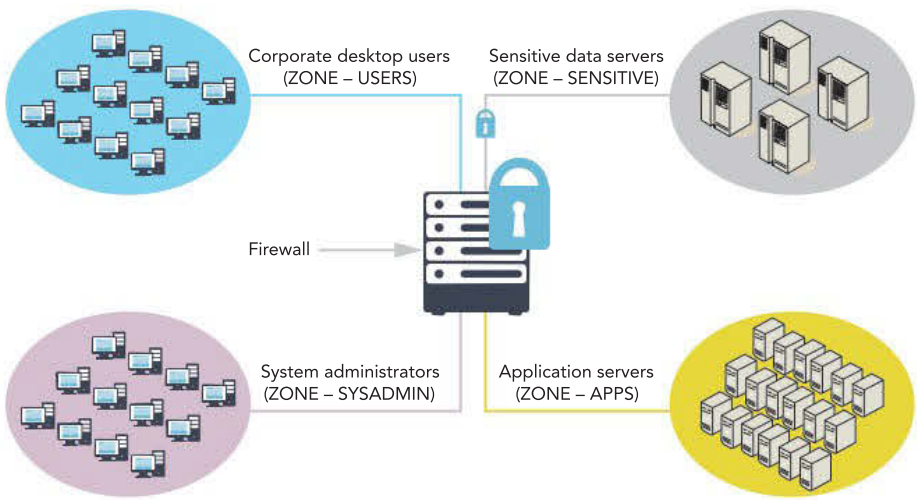


FIGURE 6.22 Zoned network areas

Barrier techniques

Barrier techniques involve using physical barriers to stop unauthorised people gaining access to hardware and software. Barrier techniques include:

- fences
- locks
- guards
- CCTV
- gates.

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

The theory behind barrier techniques is that only authorised personnel will be granted access to the location where they could access hardware devices or gain access to data and information. In some situations, a combination of barrier techniques are applied to reduce the threat of unauthorised access.

Biometrics

Biometric security is authentication that uses the unique biological characteristics of an individual to verify that they are who they are claiming to be. Types of biometric authentication include face, fingerprint, iris, signature and voice recognition (see Figure 3.31 on page 154).

Biometric authentication requires users to initially submit their biological data, which is then stored in a database. When the user attempts to access a room, building or digital device, they are required to input their biological data again. This is compared to the data previously stored in the database. If both sets of biometric data match, authentication is confirmed.

Biometric authentication is becoming far more widely used in society as the cost of the hardware and software becomes more affordable. It offers greater security than username and password but biometric software is still not 100% reliable. Systems can still throw up errors such as an authorised user fails, or an unauthorised user passes, the authentication.

Backing up

Backing up involves making a copy of the data. The copy is stored in a separate location to the original data. If the original data becomes corrupt (or is lost), then the data can be retrieved from the backup.

Similar to some other security controls, a backup does not stop threats to data and information. Rather, a backup provides protection against data loss. If the original data is corrupted or damaged, due to a data threat, the backup can be retrieved so data loss is minimised.

Many organisations follow a backup strategy when creating copies of their data and information. A **backup strategy** requires a series of elements or procedures to be identified and set out how the backup will occur.

A backup strategy may consist of the following elements.

- Timing of the backup
- Type of backup
- Storage media used
- Location of the backup
- Backup personnel

Timing of the backup

For many organisations, it may take a significant amount of time to complete a backup. It is also preferable that while a backup is being completed that the files containing the data and information are not being used. While a backup is being completed, this will use resources of the system that may affect the performance of a network or information system. For these reasons, it is often preferable to perform a backup at a time where there is either no activity, or very little activity, on the network or information system to perform the backup. Many backups are completed overnight or on weekends so there is little or no impact on the normal operations of the organisation.

9780170440875

Type of backup

There are two types of backups that could be performed: a full backup or an incremental backup.

A **full backup** involves all of the organisation's data and information to be backed-up at one time. The advantage of this is that all the data and information is backed-up at the same time, so if data is lost, the last full backup can be restored quickly and easily. A disadvantage of a full backup is that it can be very time-consuming.

To save time, many organisations perform an **incremental backup**, which involves creating a full backup from time to time (that is, weekly) and performing a partial backup in-between full backups (that is, daily). During the partial backup, only files that have changed since the previous backup are copied.

The advantage of an incremental backup is that the partial backups are less time-consuming; however, if data loss does occur, restoring the data is more complicated and time-consuming since multiple backups need to be restored.

THINK ABOUT DATA ANALYTICS 6.13

What is your backup strategy? How often do you backup your files? Is it a full backup or an incremental backup?

Storage media used

The type of storage media used to store the backup also needs to be considered when creating a backup strategy.

Hard disk drives, solid state drives, USB flash drives, portable hard drives, cloud storage or file servers are just some of the options an organisation has when developing a backup strategy.

When choosing the type of storage media to be used to store the backup, there are a number of factors to consider, including:

- the amount of data to be backed-up
- the location where the backup will be kept
- data transfer rate of the storage media
- the cost of the storage media.

THINK ABOUT DATA ANALYTICS 6.14

Where do you backup your files to? Write out a backup strategy that would suit you.

Location of the backup

There are several options when considering where to locate the storage media used to store the organisation's backup. Many organisations arrange to store the backed-up files in a separate location (that is, off-site) to the original files. This may involve the network manager taking a portable hard drive home each week, or it can involve using a cloud provider to create an online backup that is stored at a remote location.

Storing the backup on the same site as the original files allows quick access to the backup if the original files become damaged or lost. However, if the original files and backed-up files are stored in the same location, there may be a chance they are both lost to the same threat to data (for example, fire or theft). Cloud backups are becoming increasingly common.

THINK ABOUT DATA ANALYTICS 6.15

List the advantages and disadvantages of locating a backup in the cloud.

Backup personnel

The people responsible for a backup also need to be identified in the strategy. This can include the name and contact details of the staff responsible for completing the backup, staff responsible for restoring the backup and also emergency numbers in case these staff cannot be contacted when required.

9780170440875

CASE STUDY

Focused Electronics

Focused Electronics is a new but rapidly growing organisation. Focused Electronics conducts a full backup each Monday morning at 9 am. The average size of the daily backup is 1.8 TB and Dexter, the network manager, backs up the data onto a 2 TB portable hard drive. The backup takes around two days to complete using UTP cabling with data transfer speed of 100 Mbps. Once the backup is completed, Dexter leaves the portable hard drive on top of the file server, so it is easy to locate if needed.

Identify the advantages and disadvantages of Focused Electronics' backup strategy. Suggest an alternative strategy for the organisation.

THINK ABOUT DATA ANALYTICS 6.16

Some security experts recommend that organisations use cross cut shredding when disposing of paper-based data and information. Give reasons why.

Shredding documents

All hard copies no longer required by an organisation should be shredded to stop unauthorised users gaining access to the data. This ensures that data is disposed of securely.

Shredding documents involves cutting paper into small strips or pieces, making it very complex and time-consuming to reconstruct the strips to retrieve any meaningful data. Cross-cut shredders cut both ways to form confetti-like pieces. The size of the piece can be adjusted depending on the security level. Some organisations either pulp or burn the shredded material.

Check authorisation credentials

Checking authorisation credentials involve checking user's credentials when accessing restricted areas. This stops unauthorised people gaining access to where data, information or hardware is located.

This may involve:

- security guards
- swipe cards
- security fobs
- biometrics
- CCTV
- keypads
- two-factor authentication (such as SMS).

Checking authorisation techniques can often be combined with barrier techniques to assist in the protection of data and information.

Managing data on a network

A range of hardware, software and technical protocols are available to manage the data and information used by information systems and networks.

Hardware

Hardware used to manage, control and secure data includes:

- network interface card (or wireless adapter)
- server
- wireless access point
- router
- **switch**
- **modem**.

9780170440875

Network interface card (or wireless adapter)

Each device requires a network interface card (or wireless adapter) to be installed to allow access to the network or information system. A **network interface card** slots into the motherboard of a device and then provides ports to allow the device to connect to a network. A wireless adapter allows a device to send and receive wireless signals.

Many devices come with both a wired network interface card and a wireless adapter pre-installed. This allows the device to communicate with other devices.



Shutterstock.com/designelements

FIGURE 6.23 A network interface card

Server

A server is a device that is used to provide services to other devices connected to a network – and many networks have a file server. All of the files of the organisation are stored on the one device, then users using other devices (known as clients) can access the data via the file server.

Often all of the data and information related to an information system will be stored on a web server. When a user wishes to access the information system, they will be directed to the web server where they will be able to access the data and information required.

Servers can be set up to control and manage particular tasks on a network including:

- file server
- web server
- proxy server
- database server
- print server
- email server.

6.17 THINK ABOUT DATA ANALYTICS

Identify the types of servers used in your school or institution.

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

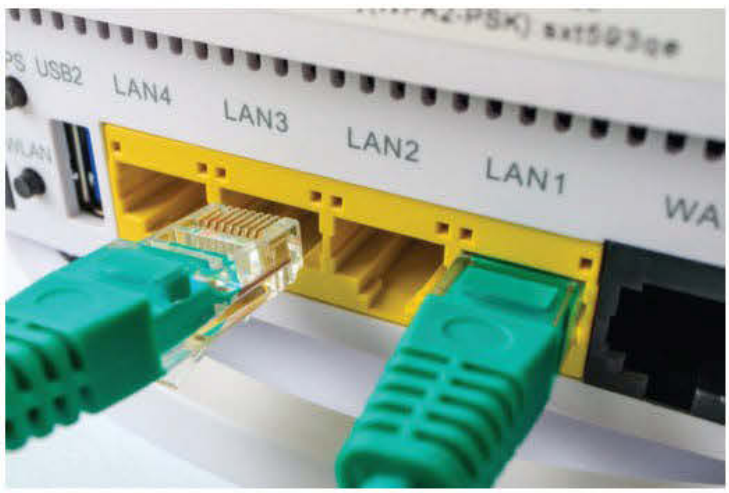
Wireless access point (WAP)

A wireless access point (WAP) allows devices to wirelessly connect to an existing wired local area network. A WAP allows the LAN to be extended to cater to more users. The WAP is connected by cable to the network. Once a data packet is received, the WAP converts the data into a wireless signal and this is transmitted over the network to the other devices. Most wireless access points have a reach of around 50 metres.

When data is sent wirelessly using a wireless access point, the 802.11 protocol is used. A WAP contains both a radio transmitter and antenna to facilitate the sending and receiving of wireless signals.

Router

A router is a device that is used to connect two separate networks together. It is often used in a home network to connect the home LAN to the internet. When data packets are received by the router, the router directs the data to the correct location. Routers are also capable of converting data using one protocol to another protocol.



Shutterstock.com/Andrei Metelev

FIGURE 6.24 A router used to connect two separate local area networks

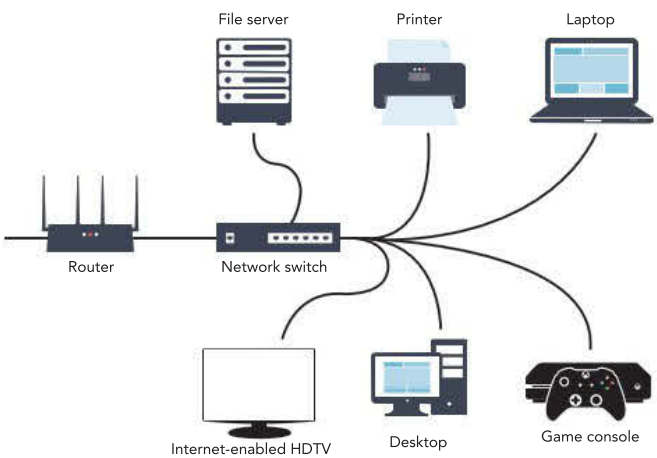


FIGURE 6.25 A switch can connect multiple devices on a LAN.

Switch

A switch is used to connect multiple devices to a network and allows them to exchange data. It is used to group together devices in a room or building. This means only one connection is needed to another switch in a different room or building.

Modem

Traditionally, a modem is used to connect a network to the internet using a telephone line to communicate data. A modem converts a digital signal (binary) into an analogue (audio) signal so data can be sent using a telephone line.

9780170440875

The use of telephone lines as a medium for data transfer was more convenient when the internet was first established since most homes and businesses already had a telephone line connection. This avoided the need to install a new dedicated connection at each location.

Devices are now available that combine the features of a modem with the features of a router. In particular, modem/routers used to connect to the NBN offer the functionality of both in the one device. Most NBN connections are fibre-to-the-node (FTTN) or fibre-to-the-curb (FTTC), so a telephone line is still used to communicate data into the house (therefore, modem functionality is still required).

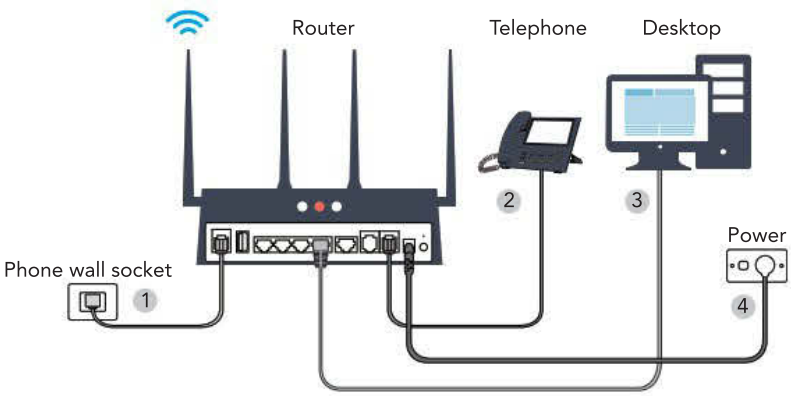


FIGURE 6.26 A modem used to connect to the NBN

THINK ABOUT DATA ANALYTICS 6.18

Is a telephone line connection essential to connect to the National Broadband Network (NBN)?

Merging technology

As technology has improved, many devices have been merged together – such as those listed here.

- Modem/router
- Router/wireless access point
- Router/switch
- Modem/wireless access point/router

Network operating system (NOS)

Software is used to control, manage and secure data, and this includes a **network operating system (NOS)**.

Just like your laptop or smartphone needs an operating system, each network needs a network operating system. Types of NOS include macOS X and Microsoft Windows Server. A NOS helps to control and manage the operation of a network including managing:

- users
- user permissions
- file access
- print jobs
- security updates.

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan | ☐ Collect complex data sets | ☐ Analysis | ☐ Folio of alternative designs | ☐ Infographic or dynamic data visualisations | ☐ Evaluation and assessment | ☐ Finalise report or visual plan

Technical protocols

Security protocols were discussed on page 280. A protocol is a set of rules or guidelines outlining how data will be communicated between devices. Network or **technical protocols** are used to ensure that devices on a network can communicate with each other.

Technical protocols include the following.

- Ethernet
- 802.11
- TCP/IP
- File transfer protocol (FTP)

Ethernet

The **Ethernet protocol** is a set of rules about how data is communicated in wired networks. On an Ethernet network, data travels in frames. Each frame is about 1.5 KB in size. Because each frame is organised in the same format, each device can understand the data received.

802.11

The 802.11 protocol is used on wireless networks. Data travels in frames (similar to the Ethernet network). The higher the letters at the end of 802.11, the faster data is transmitted (for example, 802.11n is faster than 802.11g). In 2016, version 802.11ac was released.

TCP/IP

The **Transmission Control Protocol/Internet Protocol (TCP/IP)** is a set of rules that allows communication between two networked devices.

TCP/IP is a combination of two protocols (that is, TCP and IP). It ensures that the messages travelling over the internet reach the destination IP address. TCP/IP can also be used on local area networks.

Transmission Control Protocol (TCP)

When data is sent over the internet using TCP, it can undergo a number of processes, such as:

- converting the data into packets
- sending each packet towards the destination
- arranging the packets in the correct order when they arrive
- reassembling the packets back into the correct format at the destination
- sending conformation to the sender that all of the packets have arrived.

TCP is also responsible for resending any packets lost during the communication process.

Internet Protocol (IP)

The IP is responsible for addresses and routing of the packets of data to the correct destination, utilising packet switching.

Packet switching

When data is sent across the internet, it is broken down into small packets by TCP. When each packet is directed across the internet by IP, it is possible that each packet travels a different route to the destination.

THINK ABOUT DATA ANALYTICS 6.19

How does the use of packet switching assist to increase data transfer rates across the internet?

9780170440875

As each individual packet is sent, IP identifies the most efficient route to the destination at that point in time. As the next packet is sent, then the most efficient route is recalculated. This improves the data transfer speed of sending data over the internet.

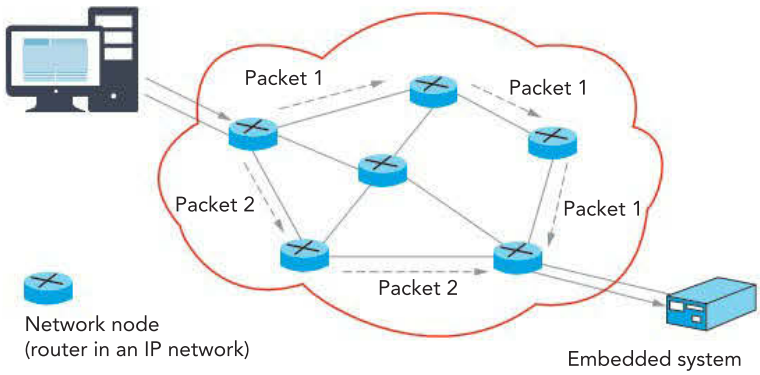


FIGURE 6.27 Packet switching

Accessing websites

A website consists of a number of web pages that are accessible over the internet by entering the site's domain name or internet address. All the resources of the website (for example, pages, files, multimedia content) is stored on a web server. Behind each domain name is an IP address that directs the request to the location of the web browser where the resources are stored.

File transfer protocol (FTP)

File transfer protocol (FTP) is a protocol used to transfer files between two devices. With FTP, a user can upload, download, delete, rename, move and copy files on a server. A user typically needs to log onto an FTP server. Once logged on, a user can manage multiple files in a quick and efficient manner.

Network attached storage and cloud computing

When considering where to store an organisation's data and information (and backups), two options to consider are network attached storage and cloud computing.

Network attached storage (NAS)

Network attached storage (NAS) is a device that offers data storage capabilities. Users of the network can save or retrieve data from the NAS. NAS in some respects is similar to, but lacks some of the functionality of, a file server.

The advantages of NAS is that it has a large storage capacity and it is easy to add further storage capacity if required. It is relatively inexpensive and it does not require a high level of expertise to set up and manage. Since the NAS is located within the LAN, an organisation has control over the data stored and who can access the data. It can also track who changed or deleted files or data. Transmission speeds are usually faster in a NAS than cloud storage

9780170440875

because access does not require an internet connection. Furthermore, a feature of a NAS is that it has the capacity to create backups within the one device using a RAID system. Because the NAS is located on-site, it is readily available if needed to restore lost data and information.

An advantage that can also be a disadvantage with NAS is that the data is still stored on-site. In the case of a disaster at the location (or user error), then there is a chance of data loss.

Cloud computing

Cloud computing involves storing data in a remote location, using an internet connection to transfer data. Often the service will be provided by a third-party organisation. The data is located on a storage device at the remote location.

Advantages of cloud computing is that it offers off-site storage, meaning that the data is secure if a disaster happens at the location of the LAN. The data can be uploaded or downloaded from the cloud from any location. A user does not need access to the LAN to access the data.

The disadvantages of cloud computing is that it requires an internet connection to access the data. If the internet connection is not working, the data and information cannot be accessed. The data transfer rate of the connection can also affect the effectiveness of using cloud storage.

The data is stored in a remote location, and often controlled by a separate organisation, so there is an element of control lost over the data. If data is required to be deleted, an organisation may never be sure if all copies have been disposed of permanently.

THINK ABOUT DATA ANALYTICS 6.20

Identify some other advantages and disadvantages of both cloud computing and network attached storage.

9780170440875

6 CHAPTER SUMMARY

Essential terms

802.11 protocol a protocol used on wireless networks; data travels in frames (similar to Ethernet)

access log activities performed on a network, information system or web server

access restriction (also referred to as permissions) limiting the user access to particular files or particular functions of an information system

accidental threat a threat with the potential to cause harm to data and information, but is unintentional and unexpected (for example, it is an accident)

asymmetric encryption using a public key and a private key to encrypt and decrypt the same message

audit trail establishing a sequence of actions in an information system

backing up making a copy of the data; the copy is stored in a separate location to the original, and if the original data becomes corrupt or is lost, the data can then be retrieved from the backup

backup strategy a series of procedures about how a backup will occur

barrier technique physical barrier to stop unauthorised people gaining access to hardware and software

biometric security devices to check the unique biological characteristics of a person

Bluetooth short-range radio waves that transmit data among bluetooth-enabled devices

brute force attack repetitive attempts at accessing an information system using a variation of possible usernames and/or passwords

checking authorisation credentials checking a user level of access when accessing restricted areas

client/server network where a central server stores data and software, and client devices request data from the server

cloud computing involves storing data in a remote location, using an internet connection to transfer data

deliberate threat a threat to data or an information system that has been created on purpose

disaster recovery plan (DRP) a strategy prepared in advance to explain how to prepare for, survive, and recover from, catastrophic data loss

encryption encoding or changing data so if the data was read by an unauthorised user they would not be able to understand the data

Ethernet protocol a set of rules about how data is communicated in wired networks

event-based threat a threat to data or an information system that is a result of a natural event; this can include fires, flood, heatwaves, storms and earthquakes

fibre-optic cabling a wired transmission media containing shards of glass that reflect pulses of light generated by small lasers or light-emitting diodes (LEDs)

firewall software that monitors all incoming (and outgoing) internet traffic to and from a LAN; a firewall examines the contents of each data packet and determines if the data packet should be allowed access to the LAN

full backup backing up all of an organisation's or individual's data and information at one time

hardware loss when a device or hardware component is lost by accident

9780170440875

hardware or software malfunction if the hardware or software stops working as intended

hardware security controls hardware-based measures that are used to protect data and information

incremental backup creating a full backup from time to time (weekly) and performing a partial backup in-between full backups (daily)

Internet of Things a network formed by 'smart devices' such as mobile phones, wearable devices, headphones, and other devices that have an 'on-off' switch to the internet and/or to other Internet of Things devices

local area network (LAN) two or more devices are connected in the same geographical area

malware malicious software that is designed to either damage, disrupt or gain unauthorised access to an information system

mobile network (also known as a cellular network) telecommunication networks that allow users to communicate using their mobile device

modem a device used to connect a network to the internet; most internet connections use a telephone line to communicate data

network consists of two or more digital system devices connected together, using transmission media

network architecture the layout of a network, including the hardware, software, protocols, and transmission media used

network attached storage (NAS) a device that offers data storage capabilities; users of the network can save or retrieve data from the NAS

network interface card a device that slots into the motherboard of a device and then provides ports to allow the device to connect to a network

network operating system (NOS) software that control and manage the operation of a network

node a connection point in a network that can create, receive, store and send data along network routes

peer-to-peer network a network where all the devices are equal; all devices store and share data with other devices

phishing sending hoax emails or messages to users pretending that the communication is from a legitimate source with the aim of extracting sensitive information

power loss any time the power to an information system is lost

power surge one type event-based threat; an abnormally high voltage of electricity received at a location in a short amount of time

private key encrypts or decrypts data by a user that has the key installed on their device

protocol a set of rules or guidelines outlining how data will be communicated between devices on a network

public key used to convert a message into an unreadable format; decryption of the message is done by a different private key

router a device used to connect two separate networks together; often used in a home network to connect the LAN to the internet

security control a measure to protect data and information from threats, either accidental, deliberate or event-based

security protocol a set of rules or guidelines for communicating data securely

server a device that is used to provide services to other devices connected to a network

shredding cutting documents into small strips or pieces

software security controls software-based measures that are used to protect data and information

standalone device any piece of computing equipment that can perform its function without the need of another device, computer or connection

switch a device used to connect multiple devices to a network

symmetric encryption requires both the sender and the receiver to use the same private key

system protection software contains a number of smaller software applications, that combined, will assist in protecting an information system or device from a range of threats (in particular, malware)

technical protocol a protocol that is used to ensure that devices on a network can communicate with each other

9780170440875

threat anything with the potential to cause harm to data and information stored and communicated between information systems

Transmission Control Protocol/Internet Protocol (TCP/IP) the protocol used to connect devices on the internet; some cabled networks also use TCP/IP rather than Ethernet; TCP/IP sends data in packets

unshielded twisted pair (UTP) used in most wired networks; wires are twisted into pairs to reduce interference

user error any error that has been caused by the user, rather than due to the hardware or software not working as expected

username and password also known as a login, a unique combination of characters/words/phrases that identifies someone on a computer system

version control management of files and software so only one version of the most up-to-date file is in existence at a time and so that major reworkings can be reversed if needed

wide area network (WAN) a network that is located in more than one geographical location

wired network a network where the devices are connected using a physical cable

wireless access point (WAP) a device that allows other devices to wirelessly connect to an existing wired network

wireless network a computer network that uses wireless data transfer between networked devices

zoned security strategies security protocol that involves breaking down a network into discrete sections or zones to limit damage caused by security threats

Important facts

1. A **standalone device** is a device that functions without the need to be connected to any other device. A **network** consists of two or more devices connected together.
2. Advantages of a network include sharing data and information, allowing communication and sharing hardware and software.
3. Types of networks include **local area networks (LANs)** and **wide area networks (WANs)**, while network architecture includes peer-to-peer networks and client/server networks.
4. A **wired network** uses physical cables to communicate data between devices, a **wireless network** can transmit data with the need of cabling, while a **mobile network** uses a telecommunications network coverage to send and receive data.
5. When using a network, there can be an increased threat to data and information. A **threat to data** can result in data being lost, damaged or stolen.
6. Types of threats to data and information systems include **accidental** (for example, user error, hardware malfunction or power loss), **deliberate** (for example, malware, phishing or data theft) or **event-based** (for example, power surge, fire or heatwave).
7. **Security controls** are measures used to reduce the chance of data loss or theft due to the threats to data.
8. **Security controls** can be either software-based or hardware-based. Examples of software-based security controls include a username and password, encryption, system protection software or firewall. Examples of hardware-based security controls include backing up, zoned barrier techniques and biometric security.
9. There is a range of hardware, software and protocols used to control and manage data and information utilised by an information system or network.
10. **Hardware** used in a network include network interface cards, servers, routers, modems and switches; a network operating system offers a range of functions to manage and control network usage.
11. **Technical protocols** outline rules about how data is transferred between devices. **Network protocols** include Ethernet, 802.11 and TCP/IP.
12. When an organisation considers options relating to storing and backing up data, two options include **network attached storage (NAS)** and **cloud computing**.
13. **Network attached storage** can be stored locally and allows the organisation full control of the data and information. **Cloud computing** offers the security of an off-site storage, but the organisation may lose some control over the management of the data stored.

9780170440875

TEST YOUR KNOWLEDGE

Review quiz

Networks

1 List three advantages of networks.
2 Explain how a local area network differs to a wide area network.
3 Discuss the difference between a peer-to-peer network and a client/server network.
4 Describe one advantage of a wired network compared to a wireless network.
5 State which wired transmission media produces faster data transfer rates. Explain how the faster transfer rate is achieved.
6 Describe one advantage of a wireless network compared to a wired network.
7 Define the 'Internet of Things'.

Threats to data and information

8 Define 'accidental threat'.
9 Describe a specific example of user error.
10 Explain how hardware malfunction differs to hardware loss.
11 State and describe one type of malware.
12 Describe the purpose of phishing.
13 Describe one example of an event-based threat.

Physical and software security controls

14 State if security controls guarantee the safety of data and information. Give a reason for your answer.
15 State the advantages of using a passphrase as a password.
16 Compare and contrast access logs and an audit trail.
17 Explain why an organisation should use both system protection software and a firewall to protect a network.
18 List the elements of procedures that should be included in a backup strategy.
19 Describe how biometric security techniques can be used to check the credentials of a user.

9780170440875

TEST YOUR KNOWLEDGE

Managing data on a network

20 Explain why a device needs a network interface card or wireless adapter.

21 Describe the role of a server on a network.

22 Explain the difference between a router and a modem.

23 List three functions of a network operating system.

24 Compare and contrast the Ethernet and 802.11 protocols.

25 Explain how TCP/IP improves data transmission rates.

Network attached storage and cloud computing

26 Describe network attached storage.

27 Describe cloud computing.

28 Explain one advantage of network attached storage compared to cloud computing.

29 Explain one disadvantage of network attached storage compared to cloud computing.

30 Discuss one concern about storing data with a third-party organisation.

31 Describe one disadvantage of cloud computing related to the disposal of data and information.

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

APPLY YOUR KNOWLEDGE

Nelson College has campuses located at three different locations. The junior school is located at one site, the middle school is located across the road from the junior school, and the senior school recently moved to a new location five kilometres away from both the junior and middle schools.

The college has installed three networks allowing staff and students in each location the ability to access files, email and the internet from any of the three locations. There are over 900 students and 100 staff at the college. A network operating system is installed to control and manage the network.

At each location, a switch in each building is connected to the network using fibre-optic cabling, while unshielded twisted pair cabling has been used to connect the switch to devices within the building. These devices include printers, projectors and wireless access points. All staff and students access the network wirelessly, using their device, via a wireless access point.

Nelson College has created a dedicated connection, under the road, between the junior and middle schools, using fibre-optic cabling, while they use an internet connection to communicate with the local area network located at the senior school.

A firewall is installed on each internet connection, and system protection software scans the network looking for malicious software. Each user requires a username and password to access the network. A log is kept detailing user activity on the network.

The college creates a backup of all data once a month. The backup is completed on a Friday night and takes 48 hours to complete. The backup is stored on a network attached storage device and this device is located in the same building as the servers of the network.

1. Explain whether the network installed at Nelson College would be considered a local or wide area network. Give a reason for your answer.
2. List two advantages of the network for staff and students of Nelson College.
3. Explain the role of a server on the network.
4. Discuss an advantage for staff and students to be able to connect to the network wirelessly.
5. Identify two security controls used to protect data and information stored and communicated on the network. For each control measure, describe the type of threat it helps to reduce.
6. Outline one other security control that Nelson College could implement.
7. Outline ways in which the college backup strategy could be improved.
8. Describe the role of a switch on the network.
9. Explain one advantage and one disadvantage of a network attached storage device.

9780170440875

CHAPTER

7 Cyber security measures

KEY KNOWLEDGE

After completing this chapter, you will be able to demonstrate knowledge of:

Data and information

- characteristics of data that has integrity, including accuracy, authenticity, correctness, reasonableness, relevance and timeliness

Interactions and impacts

- the importance of data and information to organisations
- the importance of data and information security strategies to organisations
- the impact of diminished data integrity in information systems
- key legislation that affects how organisations control the collection, storage, communication and disposal of their data and information: the *Health Records Act 2001*, the *Privacy Act 1988* and the *Privacy and Data Protection Act 2014*
- ethical issues arising from data and information security practices
- strategies for resolving legal and ethical issues between stakeholders arising from information security practices
- reasons to prepare for disaster and the scope of disaster recovery plans, including backing up, evacuation, restoration and test plans
- possible consequences for organisations that fail or violate security measures
- criteria for evaluating the effectiveness of data and information security strategies.

Reproduced from the VCE Applied Computing Study Design (2020–2023) © VCAA; used with permission.

FOR THE STUDENT

In this chapter, you will focus on data and information security and its importance to an organisation. You will also be introduced to key legal requirements for collection, storage, communication and disposal of data and information.

FOR THE TEACHER

This chapter is based on Unit 4, Area of Study 2 and, together with Chapter 6, provides the key knowledge required to complete Unit 4, Outcome 2. At the end of chapters 6 and 7, students should be able to respond to a teacher-provided case study to investigate the current data and information security strategies of an organisation, examine the threats to the security of data and information, and recommend strategies to improve current practices.



Shutterstock.com/Ekaphon maneechot

9780170440875

The importance of data and information to organisations

From data to information

Data and information are terms frequently used in information technology. You were introduced to these terms in Chapter 1. Indeed, the main purpose in using computers to solve information problems hinges on the difference between data and information. As discussed in previous chapters, the term **data** refers to raw, unorganised facts, figures and symbols fed to a computer during the input process. Data can also mean ideas or concepts before they have been refined.

Information is obtained when data is manipulated by a computer's processor into a meaningful and useful form (also becoming output). This can be achieved by organising the data and presenting it in a way that suits the needs of the intended recipient. Information is then used to assist in decision-making.

Chapter 1 explored the characteristics of data such as integrity, accuracy, authenticity, correctness, reasonableness, relevance and timeliness.

Organisational goals and information systems

Organisations tend to go through many changes over time, and these changes are generally the result of a strategic plan – a process for identifying long-term goals within an organisation. For example, a school is an organisation, and a school needs to establish a strategic plan on how it intends to maintain or increase their current student enrolment level, introduce new courses and perhaps erect a new building. This type of planning looks beyond the day-to-day running of the organisation and concentrates on future developments that could range anywhere from 2 to 25 years.

Identifying organisational goals is important to understanding how and why organisations operate.

Once an organisation has developed a strategic plan, a mission statement is developed based on the organisation's purpose, visions and values. The **mission statement** is the basis for establishing a set of common goals that will help accomplish the organisation's aims. These are known as organisational goals. See Figure 7.1 (page 299) for an example of a mission statement.

Mission statements concentrate on the present, whereas a vision statement focuses on the future. A mission statement explains the company's reason for existence. It describes the company, what it does and its overall intention. A vision statement describes the organisation as it would appear in a future successful position.

Organisational goals explain how an organisation intends to go about achieving its mission. For example, a car manufacturer might identify its mission as increasing its market share and making a profit. Establishing goals of introducing a new model of car each year and providing the highest quality spare parts to customers will enable it to achieve that mission.

A school, for example, may establish targeted goals, such as the introduction of a vocational education course within two years, or the completion of a new science and technology centre in the next five years. To achieve these goals, the organisation needs to develop a list of **objectives**. Objectives are small, achievable tasks undertaken to accomplish a big task. They have a set target to achieve; that is, they are measurable. For example, an objective may be to increase student enrolments by 10% within 12 months.

9780170440875

CASE STUDY

A mission in chocolate – Lindt Cocoa Foundation

Vision

Our vision is that the cultivation, production and processing of raw materials used in chocolate has a positive effect on sustainable agricultural development in origin countries.

Mission

Our mission is to foster sustainable agriculture in developing and emerging regions through the support of innovative projects that aim at enabling farmers to improve their practices in the cultivation, production and processing of raw materials used in chocolate.

In order to reach the mission, the foundation focuses on the following areas:

- Support projects that aim at creating the motivation, capabilities and capacity of farmers to improve their farming practices
- Support projects that aim at fostering an enabling environment and removing constraints farmers face to improve their practices
- Support research projects that aim at increasing the understanding about the farmers and their environment in order to improve the efficiency and effectiveness of projects



Shutterstock.com/ Nenov Brothers Images

FIGURE 7.1 Lindt Cocoa Foundation's vision and mission statement

Every business has a different set of organisational goals. Some have financial goals, such as making large profits, while others want to be more competitive by increasing their market share. Depending on the type of organisation, the goals will differ. For example, non-profit organisations such as charities have goals focused on providing services to members or people who may be disadvantaged in the community.

Profit-based public organisations want to provide shareholders with maximum returns in the form of dividends and growth in share prices.

Achieving the goals of an organisation can often be feasible only by purchasing or modifying an information system. However, objectives determine the size and type of information system that is needed. Depending on the organisation's goals and objectives, one or more type of information system (enterprise computing system, transaction processing system, business-support system, knowledge-management system and user-productivity system) may need to be developed.

If an organisational goal is not being reached, the organisation has a problem. Developing or improving an information system can help solve the problem and allow the organisation to achieve its goal.

Consider the needs of a simple organisation such as a sporting club. A small club might have only one or two goals. These might be as simple as keeping an accurate record of members' names and addresses and whether they have paid their subscriptions. As the club expands, it is likely that the goals and objectives will grow. Table 7.1 (page 300) illustrates how the goals and objectives of an organisation can influence the type of information system that needs to be developed.

THINK ABOUT DATA ANALYTICS 7.1

Locate mission statements from two competing companies, such as BMW and Mercedes Benz.

- How do their missions differ?
- How are their missions similar?

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

Organisational goals and objectives will often relate to improving the efficiency or effectiveness of operations.

TABLE 7.1 Common goals of an organisation

Goal	Explanation
Increase the company's profit margin	Businesses exist largely to make money. To provide value to the owners (shareholders or owner/operator), allow for further growth and the realisation of opportunities, the business needs to increase its profits.
Expand the company	As businesses want to increase their profit margins, they may find that they need to grow. They may need to employ more people and build larger premises so that their production levels meet customer demand.
Provide quality service	In particular, non-profit organisations such as charities would perceive this to be one of their most important goals. They exist to provide service to people who are disadvantaged in the community. Department stores such as Myer would also regard this as an important goal since excellent customer service is paramount to its existence.
Maintain confidentiality	Information stored about customers, products and the workings of a company needs to be protected by an organisation. Organisations need to ensure privacy and that all information is treated with confidentiality.

TABLE 7.2 The goals and objectives of an organisation can influence the type of information system that needs to be developed.

Goal	Objectives
Improve the communication of events to members	• Send letters to members and past players. • Develop a website. • Produce a regular newsletter.
Maintain accurate and detailed membership records	• Regularly check all current members' names and contact details. • Promptly add new members' details to the club's records. • Pursue overdue subscriptions. • Inform all members of subscription renewal procedures.

A systems analyst performs both the analysis and the design of a system. As well as being a good planner, a systems analyst must be a good communicator. As part of their analysis, they are responsible for talking with users and translating their needs into a design that is passed on to programmers to build.

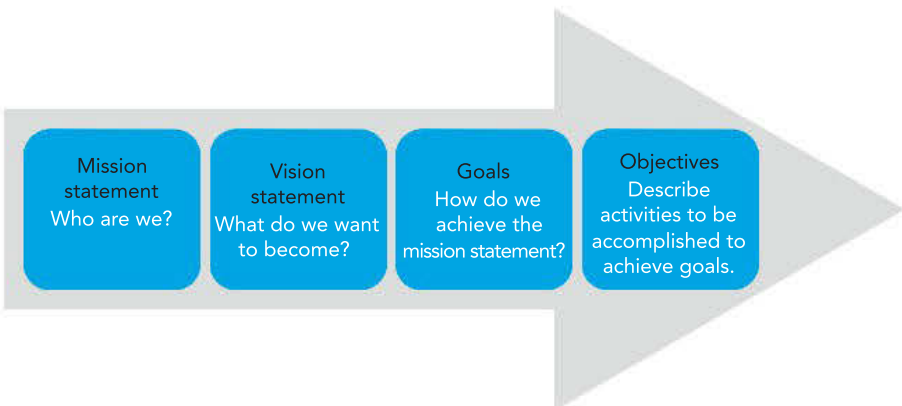


FIGURE 7.2 Relationship between mission statement, vision statement, goals and objectives

Information systems are often created to support the organisational goals. During planning stages, the systems analyst will identify a **system goal**. The system goal explains the specific role of the information system in achieving the organisational goal, and ultimately, the company's mission. Setting up the right type of information system can help an organisation make improvements in efficiency, effectiveness and decision-making.

Information management strategies

Information can be considered as a key asset in the same way as property or finance. The main function of information management is to support the business needs of an organisation by ensuring that vital information is available to support decision-making. Organisations and governments recognise this and have formulated legislation and principles to guide the use of information.

Some of the issues that arise from using information include the following.

- How do we protect information?
- How do we store information?
- How do we destroy information?
- What are the correct uses of information?
- How much information is necessary and what might be surplus?
- What information do we own and how much can be shared?

Since 1998, in Australia, legislation has been developed to manage information effectively. At the federal level, the *Privacy Act 1988* and, at state level in Victoria, the *Privacy and Data Protection Act 2014* and the *Health Records Act 2001* have been enacted. Periodically, the legislation is updated to reflect the technological advances that impact on information. Quite often, legislation is updated in a reactive manner due to new technological advances or perceived/emerging issues with existing technologies.

As organisations conclude that information should be managed as a company asset, they also recognise that managing information provides opportunities and minimises business risks. An organisation that facilitates good information management strategies has many advantages that can maximise opportunities, such as increased revenue and profitability. Advantages can also reduce costs by assisting organisations in optimising and streamlining their information assets.

Such information management strategies may also support decision-making, and identify appropriate organisational structures and governance that assist with effective information management. We can summarise these strategies into categories, such as those intended to maximise opportunities, minimise risks and to fulfil legal requirements.

The ability to find information easily can save an organisation time and resources. It can also reduce costs through removing duplicate information, which can lead to less searching time and updating of records, and removing unnecessary work.

Maximising opportunities

The large amounts of data that is being gathered and analysed is valuable to organisations. Large data sets (or 'big data') promote competition, productivity, growth and innovation. The vast amounts of data and information are equally as important as labour and capital in an organisation. Organisations can collect large amounts of detailed information on just about anything – from inventories and customer buying patterns to product preferences.

Data analysis can boost organisational performance, because it encourages organisations to make smarter decisions based on the data. Organisations can forecast in the short- and long-term and adjust their business activities as needed. This can help them to achieve their organisational goals. They can also maximise opportunities so that they can tailor products or services to their customers. For example, Woolworths collects data through its customer loyalty card, Everyday Rewards. Every time a customer scans their loyalty card during a purchase, Woolworths takes note of the items and this provides insight into customer

THINK ABOUT DATA ANALYTICS 7.2

- What are the advantages of data mining?
- What opportunities can be opened up by analysing big data?
- Why would a customer want to be tracked down based on what they buy at a particular store?
- What are the privacy issues associated with data mining?

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

information and data. All of the information collected *about* the customer is then used to create a one-to-one personalised experience *for* the customer.

Minimising risk

Threats to data and information occur every day. Data can be compromised through activities such as deliberate attacks of theft (disgruntled workers or criminals who will make money by selling the data), loss of devices (such as accidently leaving laptops at airports or in a taxi), neglect (not erasing data when recycling computer hardware) and not following appropriate data-handling procedures and policies. Collecting, storing, sending, encrypting, gathering and disposing of data is fraught with risk. To protect data and information, organisations need to manage such risks by putting measures into place, including:

- securing computers, servers and wireless networks
- utilising antivirus and anti-spyware protection, and firewalls
- storing data backups off-site and ensuring that backups occur routinely
- securing passwords
- ensuring staff are familiar with digital systems policies and procedures
- becoming familiar with legal obligations pertinent to the organisation's needs.

Data security

Chapters 3 and 6 discussed threats to data and information.

Data security has been covered in chapters 3 and 6. Refer to these chapters for a full discussion. Loss or damage to data for a business can result in an impact on reputation, trade and profit. It can also result in legal action if the *Privacy Act 1988* (see page 307) has been violated.

Protecting data involves a number of strategies, such as:

- physical security – keeping your devices safe and locked away when not in use (page 153)
- physical protection – including guards and camera (pages 281–2)
- biometric identification (pages 154–5)
- background checks of people with access to important networks and information to ascertain their spending patterns and working patterns
- security and access factors – only some personnel can access/change/save sensitive files (page 278)
- controlling access through logins requiring strong passwords that are changed regularly (pages 153–4 and 275–6)
- encryption (pages 278–9)
- use of a firewall (pages 156 and 280)
- antivirus software (pages 156–7)
- comprehensive backup strategy (pages 282–3).

9780170440875

Location of backup files

Once you have created backups, where do you put them? Ideally, your backups should be stored in a location that is safe from theft and damage caused by extremes of temperature or disasters. Most small businesses have a fireproof and waterproof safe in which valuable company documents are stored. This might also be used to store backups. It is preferable, however, to store backups at a remote location, perhaps even in the cloud. This means that if there is a large natural disaster, such as a flood or an earthquake, the backups will be safe.

One last point to remember is to ensure that backups actually work when you want to restore the data. It is important to test the effectiveness of your backup files by running a disaster recovery simulation. If files cannot be restored from the backup or the system refuses to recognise them, it is better to discover this before a real emergency.

Cloud-computing companies provide off-site storage, processing and computer resources to individuals and organisations. These companies are typically third party and they store data to remote hard drive arrays in servers that are housed in large data centres around the world. The internet provides the connection between these servers and the user's computer. One of the advantages of cloud storage includes the ability to access data from any location that has internet access, eliminating the need to carry a USB or a hard drive to retrieve and store data. The ability to share files with other people and collaborate simultaneously, such as by using Google Drive, is also an advantage. Finally, if something were to happen to the computer, such as a fire or natural disaster, and the data on it destroyed, having the data saved off-site in the cloud would prevent the data from being lost.

Google Drive allows users to upload documents, spreadsheets and presentations to Google's servers. Users can edit files using a Google application and work on them at the same time as others, so they can read or make edits simultaneously.

The importance of diminished data integrity in information systems

As described previously, data integrity refers to how trustworthy data is across its lifecycle. Organisations require data to function, and without it, they can come to a standstill. Data integrity can be compromised through user error or through deliberate acts of sabotage. Common threats that can compromise data integrity include the following.

- User error
- Data transfer errors
- Misconfiguration and security errors
- Malware and deliberate acts of theft or damage
- Damaged or lost hardware

Preserving data integrity is crucial when working with information systems. Some of the ways to do this include validating the data and input. Reducing the number of input errors increases data integrity and data accuracy. It is inevitable that errors occur when data is entered; however, there are methods used to minimise this happening. A variety of data validation techniques can be used, and appropriate error messages should inform the user what the error was and the procedure to remedy the error.

Removing duplicate or stray data from a database will ensure that sensitive data is not accessible to employees who do not have access. Backups are also necessary and prevent permanent data loss, particularly if they are targeted by deliberate acts of sabotage such as ransomware. Insider threats can also occur when an individual without appropriate

9780170440875

access privileges with an intent to cause deliberate malicious harm accesses the data and compromises it. Sometimes this threat can come from someone impersonating an inside staff member.

Whenever a breach has been reported, it is important to find the source of the breach through an audit trail to check the data integrity. Characteristically, an audit trail has the following information.

- Every action is tracked and recorded – such as creating, deleting, reading or modifying data.
- Every event is also associated with a user; therefore, there is a record of the person who accessed the data.
- Every event is time-stamped to indicate when it occurred.

Users of information systems should not have access to tamper with the audit trail. As computer systems have become ubiquitous, so too have the security risks. The widespread use of the internet, mobile computer devices, and wireless technologies has made access to data easy and this in turn has opened up opportunities to diminish data integrity, all resulting in serious financial, reputational and other damages.

Key legislation relating to data and information

There are several key laws relating to the information systems and telecommunications industries. At a federal level, the law concerned that protects the rights of creators of creative and artistic works is the *Copyright Act 1968*. Other key legislation includes the *Privacy Act 1988*, which provides protection on how information about people can be used. In Victoria, we are especially concerned with the *Privacy and Data Protection Act 2014* and the *Health Records Act 2001*. Combined, these laws govern the collection and use of private information by both government and non-government organisations at both state and federal levels. Employers and government agencies have a legal responsibility to ensure that these laws are implemented within their organisations. In addition, organisations must make employees and customers aware of their rights, as well as their responsibilities, in relation to these laws.

This section examines the key laws affecting the collection, storage, communication and disposal of data and information held by organisations.

Copyright Act 1968

The *Copyright Act 1968* is a federal law that recognises that any original creative or artistic work is the property of the person who created it. Anyone wishing to use another person's work must obtain permission and/or pay for a licence. The *Copyright Act* protects the creator of an original work from unauthorised reproduction, conversion, adaptation, transmission or publication of their intellectual property (IP), which includes:

- original literary, dramatic, musical and artistic works
- websites
- software
- electronically recorded music, films and books.

9780170440875

Copyright protection is automatic as soon as intellectual property is created and recorded in a way that can be seen or heard (for example, written, recorded, filmed or put online). The *Copyright Act* does not, however, cover ideas, concepts, styles, techniques, information, names, titles, slogans, people and images of people. You do not have to register for copyright as you do for patents or trademarks. You do not need to use a copyright symbol © or statement, but they are recommended.

You have the right to protect your own original works using technological devices, such as encryption or copy protection.

Without permission from the copyright holder, it is illegal to (with some exceptions):

- digitise a non-digital work, such as converting a DVD to an MKV file
- make or import devices or software to bypass copy protection
- remove or tamper with a copyright notice
- share copyrighted material online
- keep or share programs recorded from TV
- publish unauthorised screenshots from some web pages or software.

There are a few limited 'fair dealing' exceptions for using IP without permission: for research or study, reporting the news, criticism or review, parody or satire, disability access and professional advice from a lawyer. However, there are still constraints on these uses. It must be genuinely used for one of the purposes set out in the 'fair dealings' part of the Act and it must be 'fair' in that context. For example, for research or study purposes it is permissible to reproduce 10% of the number of pages of a printed book. There are also separate provisions for libraries and archives, governments and educational institutions. For example, if a school is copying or communicating certain copyright material for educational purposes, then they may do so provided they also pay to a copyright collecting society.

Another way of approaching exceptions to copyright is 'fair use'. This is a broad exception applicable to the United States but not to Australia. 'Fair use' can be considered more flexible than 'fair dealing' because it allows for the possibility of other uses to be considered fair even if they are not mentioned in the Act. This is useful when technological changes outpace the legislation. However, this flexibility makes 'fair use' in the United States less certain than 'fair dealing' in Australia and decisions over what can be considered fair have been reversed in the US courts.

Intellectual property might be defined as any product of human thought that is unique and not self-evident. It applies to texts (such as books and journal articles), music (both printed and recordings), videos, broadcasts and computer programs. In Australia, intellectual property is protected by the *Copyright Act 1968*. This Act was amended by the *Copyright Amendment (Digital Agenda) Act 2000*, *Copyright Amendment Act 2006* and the *Australia–United States Free Trade Agreement (AUSFTA)*.

Generally, since the *AUSFTA* was implemented on 1 January 2005, copyright applies for the life of the creator plus 70 years. The copyright holder may not necessarily be the writer of a book, musical performer, or a film director if someone else (for example, a recording company) paid for these works to be produced. Employers usually hold copyright over material that their employees create, as do film, game and music producers. Sometimes the performer may own a share in the copyright held by these organisations. Copyright can even be sold.

Before going to press, the last known amendment to the *Copyright Act* was *the Copyright Amendment (Online Infringement) Act 2018*, which relates to website blocking injunctions. Prior to that the *Copyright Amendment (Disability Access and Other Measures) Act 2017* implemented a number of amendments supported by both creators and users of copyright content. Of note – it extended the exception for exams to online exams, allowed libraries to make 'preservation copies' of 'original versions' such as manuscripts and simplified and updated the provisions that allow the making of accessible format versions for people with disabilities.

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

According to the Australian Copyright Council, 'Copyright is infringed if copyright material is used without permission, in one of the ways exclusively reserved to the copyright owner.' This means that someone may not use a whole or a part of a work, including changing or adding to it, without seeking permission from its copyright owners. For example, a student producing a video for their local sporting club must seek permission to use any music or video clips if they are not the student's own original work. Similarly, someone who imports and then sells copyrighted items from overseas without permission is considered to be in breach of copyright.

For most copyright-related criminal convictions, an individual may currently face a fine of up to $117 000 and/or up to five years imprisonment. An organisation may face a fine of up to $585 000. It is important to note that employees who infringe copyright by pirating software on work computers are liable, but their employers can also be required to share some of the liability.

The *Copyright Act* was updated to cover work published electronically when the *Copyright Amendment (Digital Agenda) Act 2000* came into effect in early 2001. Its main purpose was to extend the existing provisions of the *Copyright Act* to cover works that were produced, stored or transmitted digitally. This includes the use of web-based materials, digital sound and video recordings (including free-to-air broadcasts, CDs, DVDs and MP3s) and circumvention of technologically-based copyright protection measures. These provisions were further extended by the *AUSFTA* in 2005.

In 2006, the *Copyright Act* was further updated to provide more direction for users, such as strengthening the owner's rights to their digital material. In addition, it makes provision for users to access some legitimate copyright material without breaking the law, such as recording television and radio programs to watch or listen to at a later time (time-shifted recordings); in these instances, users are not permitted to add recordings of these recorded to their own personal library and are not allowed to distribute recordings to others.

New exemptions relating to personal use of recorded works have allowed consumers the right to make copies of works they have purchased and transfer them into other formats for personal use. This means that it is legal to copy music from CDs you own into an MP3 format to be used on a personal music player. People are also able to transfer tapes and vinyl records to an electronic format as well as convert VHS tapes to DVD.

In 2015, further amendments were made to the Act to incorporate online infringements. The amendment to the Act was intended to disrupt large-scale websites that operate outside Australia and distribute (or facilitate the distribution of) infringing material to Australian consumers. It enables copyright owners to apply to the Federal Court of Australia to block access to online locations that meet certain conditions.

The making and distributing of copies of games, music and software, even if it is not for any personal financial gain, is illegal in Australia. These acts are commonly referred to as 'piracy'.

Further information on Australian copyright laws, including information sheets on a wide variety of copyright-related questions, can be obtained from the Australian Copyright Council.

How does copyright apply to music, computer games and computer software?

In light of what you have read above, and the impact of the *AUSFTA*, you are highly restricted in what you can copy or reproduce in other formats. In terms of music, when you buy a CD you can rip a copy to MP3 format to play on your personal music player. You are permitted to download music from the internet via peer-to-peer transfers, but only if you have the permission of the copyright holder. There may be specific terms and conditions on music that you download from online music stores that allow you to make a certain number of copies for personal use, but these vary between distributors.

Computer software is treated as a 'literary work' under the *Copyright Act*. Computer games fall into several categories because they incorporate the program code plus audio and video works that may themselves be licensed from other copyright holders. Generally, you are permitted to make a backup copy of the game itself, but not any of the artistic works that may also be on the media, such as video or audio, without seeking the permission of the relevant copyright holders. You are not, however, permitted to bypass copy protection features in order to make your backup.

Under the *AUSFTA*, any device specifically designed to bypass copyright protection measures is considered illegal. You may lend a legitimate copy of a game to someone to play, but it is illegal to play an infringing copy. If your original media is destroyed, the *Copyright Act* allows you to make another backup from the first backup. Naturally, it is illegal to make multiple backup copies and distribute those to other people. The specific software licence will tell you how many times you are allowed to have the program installed simultaneously.

Copyright and cloud computing

Many copyright owners use streaming services such as iTunes, Netflix and Stan to deliver copyrighted material to consumers. These subscription or pay-per-use services provide on-demand access to large libraries of legally licensed music, films, books and other content.

Conversely, individuals tend to engage with cloud-computing services to store copyright material they have copied or 'ripped' themselves such as music files copied from a CD. The advantages of storing these copies on remote servers means content from multiple computers and devices, including mobile devices, can be accessed easily. The issue with this is that the copyright holders of the material may object to this use.

Penalties for infringing copyright

We now know that the infringement of copyright includes activities such as selling or playing pirated software, games and music, decrypting television broadcasts, removing copyright protection and importing copyright material without authorisation. Most of these actions can be tried in court as civil actions. In general, copyright infringements that involve some kind of commercial dealing are criminal offences. For civil actions, the damages vary depending on the level of infringement and compensation.

THINK ABOUT DATA ANALYTICS 7.3

- Why have personal music players (such as iPhones), caused a rethink on Australian copyright laws?
- Research 'fair use' and explain the difference to 'fair dealing'.
- What do you and your classmates see as 'fair use' for music purchased online or in a shop?
- What are the arguments for and against a US-style 'fair use' amendment to the *Copyright Act*?

Privacy Act 1988

Originally, the *Privacy Act 1988* only dealt with the handling of data by government agencies. Many people criticised these limitations because it seemed that private organisations were not required to apply even the most basic of safeguards on data they collected. Even worse, there were no regulations preventing non-government organisations from collecting data by any method and using it for any purpose without the consent of the people whose private details it concerned. In particular, the rapid growth of electronic transactions, especially over the internet, led many people to demand some sort of legal protection from those who might gather data about internet browsing habits. The government was keen to encourage the development of electronic commerce, while protecting the confidentiality of consumers and increasing public confidence in electronic transactions. These amendments have now been incorporated into the *Privacy Act 1988* and are the most significant changes to privacy laws since the inception of the legislation.

There have been several additional powers included within this Act since 1988, but its essential purpose remains unchanged. The *Privacy Act 1988* was amended by the

9780170440875

Privacy Amendment (Enhancing Privacy Protection) Bill 2012, which came into effect on 12 March 2014.

> 'These are the most significant changes to privacy laws in over 25 years and affect a large section of the community. The world has changed remarkably since the late 1980s when the *Privacy Act* was first introduced, and so the changes were required to bring our laws up-to-date with contemporary information handling practices, including global data flows,' said Australian Privacy Commissioner Timothy Pilgrim at the time.
>
> Office of the Australian Information Commissioner website — www.oaic.gov.au

What is included in the *Privacy Act*?

The *Privacy Act* includes the following.

- Thirteen Australian Privacy Principles (APPs) that apply to the handling of personal information by most Australian and Norfolk Island government agencies and some private sector organisations
- Credit reporting provisions that apply to the handling of credit-related personal information that credit providers are permitted to disclose to credit reporting bodies for inclusion on individuals' credit reports
- The collection, storage, use, disclosure, security and disposal of individuals' tax file numbers
- The handling of health information for health and medical research purposes in certain circumstances, where researchers are unable to seek individual consent
- The Information Commissioner to approve and register enforceable APP codes that have been developed
- Providing a small business operator, who would otherwise not be subject to the Australian Privacy Principles, to opt-in to being covered by the APPs

Who is covered under the *Privacy Act*?

For an individual, the *Privacy Act* provides more control over the way their personal information is handled. The *Privacy Act* allows individuals to:

- know why personal information is being collected and how it will be used and who it will be disclosed to
- have the option of not being identified or the use of a pseudonym in certain circumstances
- ask for access to personal information (including health information)
- discontinue receiving unwanted direct marketing
- ask for personal information that is incorrect to be corrected
- make a complaint about an entity covered by the *Privacy Act* if personal information has been mishandled.

Australian Privacy Principles (APPs)

As part of the *Privacy Act*, the Australian Privacy Principles were devised to set out the standards, rights and obligations for collecting, handling, holding, accessing, using, disclosing and correcting personal information.

The APPs generally apply to federal government agencies. They do not apply to local councils, or state or territory governments. Some states have their own privacy laws, such as Victoria's *Privacy and Data Protection Act 2014*.

The APPs oversee the handling of personal information by:

- Australian and Norfolk Island government agencies
- all private health service providers
- businesses that have an annual turnover of $3 million or those that trade personal information.

TABLE 7.3 The Australian Privacy Principles (APPs)

APP 1	**Open and transparent management of personal information** Ensures that APP entities manage personal information in an open and transparent way. This includes having a clearly expressed and up-to-date APP privacy policy.
APP 2	**Anonymity and pseudonymity** Requires APP entities to give individuals the option of not identifying themselves, or of using a pseudonym. Limited exceptions apply.
APP 3	**Collection of solicited personal information** Outlines when an APP entity can collect personal information that is solicited. It applies higher standards to the collection of 'sensitive' information.
APP 4	**Dealing with unsolicited personal information** Outlines how APP entities must deal with unsolicited personal information.
APP 5	**Notification of the collection of personal information** Outlines when and in what circumstances an APP entity that collects personal information must notify an individual of certain matters.
APP 6	**Use or disclosure of personal information** Outlines the circumstances in which an APP entity may use or disclose personal information that it holds.
APP 7	**Direct marketing** An organisation may only use or disclose personal information for direct marketing purposes if certain conditions are met.
APP 8	**Cross-border disclosure of personal information** Outlines the steps an APP entity must take to protect personal information before it is disclosed overseas.
APP 9	**Adoption, use or disclosure of government related identifiers** Outlines the limited circumstances when an organisation may adopt a government related identifier of an individual as its own identifier, or use or disclose a government related identifier of an individual.
APP 10	**Quality of personal information** An APP entity must take reasonable steps to ensure the personal information it collects is accurate, up-to-date and complete. An entity must also take reasonable steps to ensure the personal information it uses or discloses is accurate, up-to-date, complete and relevant with regard to the purpose of the use or disclosure.
APP 11	**Security of personal information** An APP entity must take reasonable steps to protect personal information it holds from misuse, interference and loss, and from unauthorised access, modification or disclosure. An entity has obligations to destroy or de-identify personal information in certain circumstances.

9780170440875

APP 12	**Access to personal information** Outlines an APP entity's obligations when an individual requests to be given access to personal information held about them by the entity. This includes a requirement to provide access unless a specific exception applies.
APP 13	**Correction of personal information** Outlines an APP entity's obligations in relation to correcting the personal information it holds about individuals.

Office of the Australian Information Commissioner website — www.oaic.gov.au

Credit reporting provisions

There have been changes to the credit-reporting provisions of the *Privacy Act 1988* and how credit-related personal information is collected. The *Privacy Act* also encompasses a code of practice for credit reporting. The credit-reporting provisions for consumer credit include the simplification of the language used in reports and improved privacy protections. Additionally, the process to lodge a complaint has been simplified.

Application of the Privacy Act

The *Privacy Act* applies to both electronic and manual or conventional forms of data gathering and handling by private organisations. The Act also has provisions specifically addressing the use of personal data for direct marketing via email, which can only be used with the consent of the individual concerned. It also extends to general privacy issues regarding workplace email. The Act encompasses businesses with an annual turnover of $3 million, all private health services that store health records, businesses that trade in personal information and those organisations that choose to opt-in.

Individuals also have rights under the Act, which makes for provisions on how their personal information is collected. The Act defines personal information as being:

> … information or an opinion about an identified individual, or an individual who is reasonably identifiable: whether the information or opinion is true or not; and whether the information or opinion is recorded in a material form or not.

Office of the Australian Information Commissioner website — www.oaic.gov.au

Incorrect disposal of digital systems equipment and data storage media is a major source of privacy breach.

The amended Act defines personal information as including an individual's:

- name and address
- signature
- telephone number
- date of birth
- medical records and health information
- bank account details
- photos and videos
- biometric and genetic information
- philosophical beliefs

9780170440875

- likes and dislikes
- opinions or commentary about a person
- racial or ethnic origin
- memberships of political associations
- professional or trade associations or trade unions
- religious beliefs or affiliations
- criminal record
- sexual orientation or practices.

Since the inception of the updated *Privacy Act 1988*, organisations have had to consider and review how they handle customer information by updating their technologies and their security processes to ensure they comply with the new legislation.

The Act prescribes severe penalties for serious and repeated interferences with privacy that can result in criminal prosecution and/or fines of up to $340000 for individuals and $1700000 for public and private organisations. The amount will only change if the Act is changed. In Victoria, however, the fines are variable and tied to penalty units with an amount that is adjusted on 1 July each year.

THINK ABOUT DATA ANALYTICS 7.4

Many organisations have a privacy policy listed on their website. Find out what is covered by your school's privacy policy. What information might the school have about students that should not be made publicly accessible?

Penalty units define the amount that needs to be paid for offences in Victoria. Generally, the legislation does not specify the monetary amount, but does, however, specify the penalty unit. Each year, the penalty unit is specified, such as from 1 July 2018 to 30 June 2019, one penalty unit is worth $161.19. The rate for penalty units is indexed each financial year so that it is raised in line with inflation. Changes to the value of a penalty unit take effect on 1 July each year.

Privacy and Data Protection Act 2014

The *Privacy and Data Protection Act 2014* (*PDPA*) was introduced by the Victorian Government. It replaced the *Information Privacy Act 2000* and the *Commissioner for Law Enforcement Security Act 2005*. The *PDPA* is intended to strengthen the protection of personal information and other data held by Victorian government agencies including local councils and contractors working for the State Government.

Under the *PDPA*, there is a single privacy and data protection framework. The *PDPA* uses its own Information Privacy Principles (IPPs) and organisations are obliged to act in accordance with the IPPs. As a result of the *PDPA*, a Privacy and Data Protection Commissioner has been established.

Information Privacy Principles (IPPs)

As discussed in the previous section, the amendments to the *Privacy Act 1988* in 2014 introduced new Australian Privacy Principles. It was anticipated that the new Act would replace the current Victorian Information Privacy Principles with the new principles based on the APPs. However, this has not happened, so the Victorian *Privacy and Data Protection Act 2014* continues to use the IPPs.

These APPs replaced the two sets of principles, which, since 2001, have applied to Commonwealth public sector and private sector organisations. They were known as the Commonwealth Information Privacy Principles and the National Privacy Principles.

TABLE 7.4 The Information Privacy Principles (IPPs)

IPP 1	Collection of personal information	When an organisation collects information, it should only collect the information it needs. The organisation should inform people that their information is being collected.
IPP 2	Use and disclosure of personal information	When an organisation uses and discloses personal information it is only for the purpose that it was collected for, or for secondary purpose that you would reasonably expect.

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan | ☐ Collect complex data sets | ☐ Analysis | ☐ Folio of alternative designs | ☐ Infographic or dynamic data visualisations | ☐ Evaluation and assessment | ☐ Finalise report or visual plan

9780170440875

IPP 3	Data quality	Ensure that the information collected is accurate, complete and up-to-date.
IPP 4	Data security	Information must be protected from misuse, loss, unauthorised access, modification or disclosure. Reasonable steps must be taken to destroy or de-identify personal information that is no longer needed.
IPP 5	Openness	The organisation needs to be transparent about what it does with information. Non-compliance will result in a maximum penalty for a body corporate of 3000 penalty units and 600 penalty units for an individual.
IPP 6	Access and correction	When an organisation collects information, it should allow people to see the information it collects about them and provide them with the opportunity to correct it if it is inaccurate.
IPP 7	Unique identifiers	Use of unique identifiers, usually a number, is only allowed where an organisation can demonstrate that the assignment is necessary to carry out its functions efficiently.
IPP 8	Anonymity	Where possible, people supplying information should be given the option of not identifying themselves.
IPP 9	Transborder data flows	If your personal information travels outside Victoria, your privacy protections must travel with it.
IPP 10	Sensitive information	Organisations needs to ensure that they do not collect sensitive information about people, such as their religion, political views or criminal record, without checking the applicable laws.

© Office of the Victorian Information Commissioner

THINK ABOUT DATA ANALYTICS 7.5

Stevie is a student support officer at a Victorian government school. He has access to student and parent personal details. He has been approached by an external organisation to trade these details in exchange for new computer equipment for the school.

- Identify key legislation that Stevie should consider before providing the information to the external company.
- What are Stevie's ethical responsibilities to the students, parents and the school?

Health Records Act 2001

The Victorian *Health Records Act 2001* was created to provide direction with the collection and handling of health information in both the public and private sector. It is anticipated that patients will use both private and public health services at various stages of their life. The *Health Records Act* allows people to access their own medical information, as well as establishing the health record privacy principles for both public and private medical services. The *Health Records Act* established 11 Health Privacy Principles (HPPs) to provide rights to both living and deceased people. These principles apply to the collection, use and storage of personal health information in Victoria.

The *Health Records Act 2001* applies to a deceased individual who has been dead for 30 years or less.

The Act protects the confidentiality of patients' healthcare information by allowing the information to be used only for the primary purpose for which it was gathered. This means that information about medical test results and your medical history may be used by your doctor, the hospital and any other health professionals only for the purpose of your immediate or ongoing care. This information would not be disclosed to a third party for a 'secondary' purpose (for example, your health insurance company or another hospital) without your consent. Health information may, however, be provided to third parties without your consent under certain, and strictly limited, circumstances that include requests by family members in an emergency when you cannot give your consent and your life is threatened, where there is a serious threat to public health and welfare, research in the public interest, investigation of unlawful activity and as part of a legal claim.

9780170440875

An individual who believes that the *Health Records Act* has been breached can make a complaint to the Health Services Commissioner, who will try to achieve a resolution by discussion between the parties. If a satisfactory resolution cannot be reached, the Commissioner may then serve a compliance notice on the organisation that has breached the Act to inform the organisation which area of the Act has been breached and that it must correct its procedures. The maximum penalty for an organisation is currently 3000 penalty units and 600 penalty units for non-corporate cases.

From the age of 16, teenagers can consent to medical and dental treatment with the same authority as an adult. Therefore, teenagers can see a doctor by themselves without their parents. At 18 years of age, they have the legal capacity to give consent to, and refuse, medical treatment.

Health Privacy Principles (HPPs)

Table 7.5 sets out a summary version of the Health Privacy Principles. This does not set out the full Principles and is intended for quick reference only. The principles in full can be found in the *Health Records Act 2001*.

The *Health Records Act* appointed the Victorian Health Services Commissioner to perform a range of functions administered under the Act, including conciliation, investigation and resolution of complaints.

TABLE 7.5 A summary of the Health Privacy Principles

HPP 1	Collection	Only collect health information if necessary for the performance of a function or activity and with consent (or if it falls within HPP 1). Notify individuals about what you do with the information and that they can gain access to it.
HPP 2	Use and disclosure	Only use or disclose health information for the primary purpose for which it was collected or a directly related secondary purpose the person would reasonably expect. Otherwise, you generally need consent.
HPP 3	Data quality	Take reasonable steps to ensure health information you hold is accurate, complete, up-to-date and relevant to the functions you perform.
HPP 4	Data security and retention	Safeguard the health information you hold against misuse, loss, unauthorised access and modification. Only destroy or delete health information in accordance with HPP 4.
HPP 5	Openness	Document clearly expresses policies on your management of health information and make this statement available to anyone who asks for it.
HPP 6	Access and correction	Individuals have a right to seek access to health information held about them in the private sector, and to correct it if it is inaccurate, incomplete, misleading or not up-to-date.*
HPP 7	Identifiers	Only assign a number to identify a person if the assignment is reasonably necessary to carry out your functions efficiently.
HPP 8	Anonymity	Give individuals the option of not identifying themselves when entering transactions with organisations where this is lawful and practicable.
HPP 9	Transborder data flows	Only transfer health information outside Victoria if the organisation receiving it is subject to laws substantially similar to the HPPs.
HPP 10	Transfer/closure of practice health service provider	If you are a health service provider, and your business or practice is being sold, transferred or closed down, without you continuing to provide services, you must give notice of the transfer or closure to past service users.
HPP 11	Making information available to another health service provider	If you are a health service provider, you must make health information relating to an individual available to another health service provider if requested by the individual.

* In the public sector, individuals already have this right under freedom of information.

Department of Health & Human Services, Victoria

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

Storing health records in the cloud

The advantages of storing health records in the cloud include the availability of healthcare professionals having access to health records and data to assist with accurate patient diagnosis and medications. As the population ages and grows, this carries increased expectations of living longer, and therefore storing records in the cloud provides advantages to the healthcare industry. Sharing capabilities of the technologies enables accurate analysis of medical histories to assist with:

- diagnosis
- minimising duplication and unnecessary testing
- long-term monitoring of chronic diseases.

The improved sharing of records enables medical professionals to communicate and collaborate and offer a team approach to looking after the patient.

However, data breaches are a major threat to cloud computing. Given that health records are comprehensive and highly sensitive, these are a desirable target, especially for identity fraud, theft and blackmail. Other threats include loss of data, denial of service attacks and cyber-attacks.

While there is protection for storing health records, the complexity arises from legislation from both the Commonwealth and State level. For example, Commonwealth government agencies and private sector health service providers must comply with the APPs contained in the *Privacy Act 1988*. The *Privacy Act* recognises health information as a form of 'sensitive information'. However, state-based public sector agencies must comply with state or territory-specific legislation regarding privacy, confidentiality and data management. In this case, the *Health Records Act*. Confusingly, laws vary between states and territories and there is also significant overlap between the federal and state/territory laws.

The My Health Record system contains an outline summary of individual's health information, including the medications they are currently prescribed, allergies they have, and any treatments they have received. The *My Health Records Act 2012*, *My Health Records Rule 2016* and *My Health Records Regulation 2012* provide the legislative basis for the My Health Record system.

Ethics and security practices

Despite the various laws designed to protect our information privacy, there are still many activities that people can perform that, while not illegal, are certainly questionable or morally frowned upon. **Ethics** refers to behaving in ways that are based on our morals and accepted standards. These standards may be common in a particular society or specific to a single organisation. They apply to questionable activities over and above any legal requirements. Ethics often provide us with a set of guidelines of appropriate behaviour. If we choose to ignore these guidelines, we may not be committing a crime, but we may be terminated by an employer or shunned by society.

Everyone wants the benefits of digital systems; however, intended and unintended negative effects can impinge upon people's rights. As a result, those who design, control and use digital systems have a responsibility to consider the real and potential negative effects and to eliminate or lessen them as much as possible. Sometimes even this may not be enough to justify the proposed collection, creation or storage of data and information. It is important to take into account legal objections and ethical considerations when creating, acquiring and storing data and information. The purpose for collection needs to be clear. This also needs to be articulated in the participant information statements and the consent forms provided to the people from whom information will be sought. It should also be communicated that storage of data is set up in a way so that data and information cannot be accessed by unauthorised users, and for how long a period of time the data will be held.

A **dilemma** occurs when there is a choice between two options of equal desirability. For example, if doctors are requested to send patient medical details to a central authority for a nationwide health study, parents of young children are placed in a dilemma in which they must weigh up their child's right to privacy against the benefits to society.

The standards or guidelines that determine whether an action is good or bad are known as ethics. Ethics are the moral guidelines that govern, among other things, the use of data collection. Often ethical principles/guidelines have an accompanying law, but the ethical principle is usually broader and the law applies only to certain circumstances or applications of the principle.

For example, it is ethical to obtain permission to publish photos of people on websites or in promotional material. Sometimes people may object to their images being used for these purposes. The purpose for taking the photo and how it is intended to be used need to be made clear. Ethically it is wrong to use a photo for a different purpose from that for which it was originally collected. Similarly, when using data-collection tools such as surveys, interviews and questionnaires, it is important to reassure participants that the data provided, within the limits of the law, will remain anonymous and that their individual comments will not be able to be identified by others. It is not simply that it is important to put participants' minds at rest regarding their concerns about protecting privacy; it is also important to ensure that their privacy is, in fact, protected – and also to ensure that non-participants in the larger group of which the sample is supposed to be representative are not put at unacceptable risk of suffering as a result of mistaken identification.

This may be of particular concern when deciding what must be removed to de-identify data sufficiently to protect all its potential users. It is also important when reporting personal information anonymously or using pseudonyms in a newspaper report, for example.

Workplace responsibilities

Within any organisation, employees and employers have certain responsibilities towards one another as a duty of care, as well as to the customer or client. In particular, the employer must pay staff for the work they carry out and provide a suitable work environment in which that work can take place. In return, the employee is expected to work in the interests of the organisation for the duration of time they are being paid. In addition, the organisation is expected to provide good-quality products or high levels of service to customers. If an employee is not working well for the organisation – perhaps they are rude to customers – it is understood that the manager has the right to dismiss the employee. The employee is not breaking a law, but they are going against the standards of service insisted upon by the particular company for which they work.



Shutterstock.com/Iakov Filimonov

FIGURE 7.4 An employee who goes against the standards of service of a company may be dismissed from their job.

THINK ABOUT DATA ANALYTICS 7.6

Identify five key behaviour and work standards expected of an IT professional.

The due diligence process needs to include evidence of meetings between the employer and the worker to discuss improvements needed to the worker's performance and to document goals and strategies to redress the problem. The employer needs to show that they have tried to help the worker to fix any failings.

Codes of conduct and computer-use policies

A **code of conduct** is a set of conventional principles and expectations that is considered binding on any member of a particular group. A code of conduct for a lawyer would apply to the work they do on behalf of a client, while a doctor prescribes treatment in the interests of a patient. However, it may seem that a computer programmer's position is unclear: is the

Although an employer can dismiss a worker who is operating at below the required standard, the worker can access unfair dismissal laws. If this happens, the employer needs to be able to demonstrate that due diligence has been followed.

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

programmer working for their manager or the client? A code of conduct might say they are responsible to both, even though it may seem harder to relate workplace responsibilities in the information systems sense because conventional roles seem unclear.

What happens when there is a conflict? For example, the programmer might know that there is a bug in a piece of software that they are building. This will have an impact on the accuracy of the software's data manipulation. Due to a strict deadline, the manager tells the programmer to ignore the bug, because it is more important to have the package delivered on time. In this situation, it is difficult for the programmer to know if they owe a greater obligation to their company, to the client, or to the general public who will be the end-users.

The Australian Computer Society has a code of professional conduct for its members. This code includes all aspects of behaviour and work standards expected of an industry professional. A serious breach of a code of conduct may result in the termination of an employee.

All organisations should have their own company-wide computer-use policy. This policy would explain clearly to staff what management believes should and should not be done on the computers or peripheral equipment. Employees will then be in no doubt as to what is permitted, and when. Some companies may allow employees to use company computers for their own purposes, such as network games, personal email or general internet use in their own time. Other companies may not allow employees to use computers for any purpose other than work, even during a lunch break. Students, too, are often required to follow school-based codes of conduct. For example, students may be allowed access to the internet while at school, but they may have to sign an undertaking that they will not use it for inappropriate purposes, such as downloading pornography, using chat rooms or playing online games.

Employee monitoring

If codes of conduct and computer-use policies exist in an organisation, it is in the company's interest to enforce them. Often, managers will use monitoring systems to check on what their employees are up to. Although these systems are not illegal, depending on the extent of their intrusiveness, they may be considered unethical because they invade the employee's privacy. It might be considered acceptable for supervisors to monitor some telephone calls in a call centre to ensure quality of service, but it is unethical to monitor and then possibly restrict employees' toilet breaks to certain times of the day.

Employee monitoring advantages:
- It ensures that employees are doing company work.
- It ensures that employees maintain target performance levels.
- It saves time and money.

Employee monitoring disadvantages:
- It is intrusive and may impact on employee privacy.
- Mistrust may develop between employee and employer.

The most common monitoring is of email, PC use, and web browsing. Email messages can be automatically redirected to a manager if they contain certain words or attachments, especially .jpg or .exe, or if they come from an unrecognised email address.

Computer monitoring allows a supervisor to see what an employee is keying, as well as the files on their computer and even the desktop layout. The employer may believe it is important to be able to check that employees are doing what they are paid to do, especially with regard to data entry operators who are required to perform to a high standard of keystrokes per hour. If staff know the boss may be watching what they are doing on the computer, they will be less likely to slack off or do things that are not work related.

In theory, this may save time and money for the employer. However, employees may argue that the employer is being intrusive. In extreme cases, employees may see words start to change in a report they are writing if managers make changes to their files. They may feel an increase in stress levels because they may not know when they are being 'spied on' electronically. Computer monitoring may create mistrust between supervisors and staff. While monitoring is not illegal, in some cases it may be seen as unethical because employees would consider that they are entitled to some privacy while at work.

THINK ABOUT DATA ANALYTICS 7.7

Locate your school's computer-use policy.
- What key ethical areas does it cover?
- How effective is the policy in your opinion?

Although the internet may be invaluable to an organisation for e-commerce, research or as part of a corporate intranet, managers are also wary of its misuse for personal purposes. It is difficult to restrict the sites that users may access without severely limiting the access of other internet users in the organisation. Web browsing is, therefore, usually monitored by way of automatic logs. These are stored either on the user's own machine in the form of **cookies** or on a history of sites. The server will automatically maintain a list of every site and file accessed, cross-referenced to the relevant username and the date and time. While the log file may be huge, it is very easy to search it for key words or phrases.

Again, internet usage within an organisation might be regulated by clauses of the computer-use policy or, in the case of many schools, an internet acceptable-use policy.

The organisation may not be too concerned about an employee looking up destinations for an upcoming holiday during their lunch break, but the manager may become upset if the employee looks up pornography or gambling-related sites, or even job advertisements. From the employer's point of view, it is considered unethical to be 'misusing' the internet connection, even if it is outside working hours. Similarly, employees might expect some leeway if the alleged misuse takes place out of necessity. For example, if you were planning to book airline tickets at a discount rate, but they were only available through the internet at a particular time, then your supervisor might allow you to do so.

Resolving legal and ethical tensions

Working in an organisation with ethical, legal or social tensions can be uncomfortable. Often tensions arise through the lack of clarity in policy documents that exist within the organisation to govern the use of data and information. For example, a situation might arise in which an employee needs to perform a task on the network that their manager must approve. The manager is sick, and the employee had decided to take it upon themselves to access the manager's account in order to get the task done. Although this action is not illegal, it is unethical, and it could result in disciplinary actions for the employee.

In addition to legislation such as the *Copyright Act 1968* and the *Privacy Act 1988*, organisations might have policy documents that mandate who has access to areas of the information system and the penalties for misuse of information within an organisation. The suggested six steps for handling ethical dilemmas act as a framework that can be used within an organisation to solve legal, ethical or social tension. Once the organisation has solved its problem, it may wish to update any policies concerned to ensure that the process for resolution is clearer.

Six steps to solving ethical dilemmas, as follows.

1. **Identify the problem:** What decision has to be made and what facts are required?
2. **Identify the stakeholders:** Who are they? What interests do they have? Who are the key players?
3. **Identify possible consequences:** What options are available? What are the likely consequences?

4 **Identify ethical standards:** Are there any applicable laws? Are there any codes of conduct or standards that could be applied? Is there a precedent?

5 **Evaluate options:** Identify strengths and weaknesses. Identify the option that causes the least harm. Can the decision be reversed?

6 **Make a decision:** Select the preferred option. Justify the option.

Once the organisation has solved its problem, it may wish to update any policies concerned to ensure that the process for resolution is clearer. The final decision must also be communicated to all stakeholders.

Reasons to prepare for disaster

Disaster recovery plans (DRP)

A **disaster recovery plan (DRP)** is a comprehensive scheme that explains how to prepare for, survive and recover from a disaster that affects information systems infrastructure. It is prepared in advance, tested regularly and executed immediately after a disaster to minimise disruption and restore normal business operation as soon as possible.

A well-designed DRP has several benefits.

- It provides a sense of security and confidence to workers.
- It prevents further problems that might be caused by panic.
- It reduces the amount of uncertainty and decision-making in stressful moments.
- It neatly explains procedures to new employees.

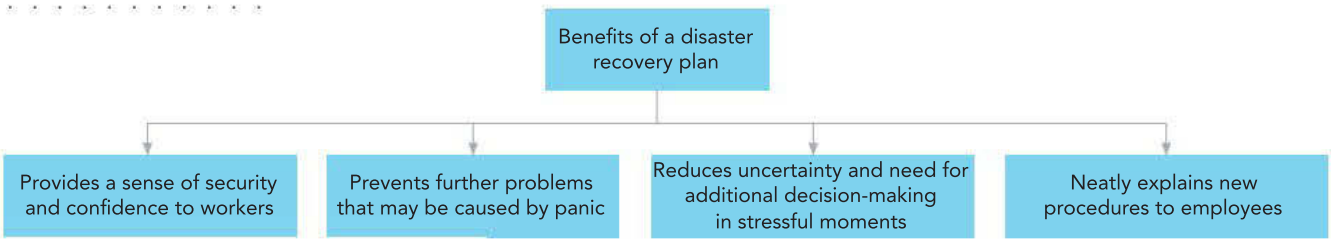


FIGURE 7.5 Benefits of a disaster recovery plan

The four main components of a DRP cover:

1 evacuation
2 backing up data
3 data restoration
4 testing of disaster plans.

Regardless of how well you protect your data and information against threats, there is still a chance that it can be damaged, lost, stolen, copied or destroyed. Only a fool would consider their data completely and permanently invulnerable. New threats are invented or discovered every day, and even the cleverest system manager cannot hope to prevent disasters caused by all known and unknown threats.

9780170440875

Disaster planning is about accepting the fact that one day, in spite of all your careful planning, your data and information will be stolen or destroyed. Whether it happens deliberately or accidentally no longer matters: the DRP is your last chance to survive and recover. Without a means of recovering lost data, an organisation struck by data tragedy faces a high probability of corporate death.

Make sure you create your DRP when you create your organisation. If you join an organisation and it does not have a DRP, make it a priority. A DRP needs to exist before disaster strikes. Inventing the comprehensive and efficient action plan while the office burns down is a bit like learning to swim after you have fallen overboard: you will have left things much too late.

Creating an effective, complete DRP requires considerable planning, and this planning must pass through distinct stages.

Scope of the disaster plan

Risk assessment

Risk assessment analyses the probability of different risks to your data and information and the potential consequences of each of those risks. For example, rural organisations would be concerned with the threat of bushfires. City organisations would be more concerned with burglary. It is more sensible to pay the most attention to the most likely and most damaging risks to your data.

- An organisation with many unskilled workers may need to focus on continual backups and staff training.
- A school may focus on vandalism and the misuse of accounts by students.
- Military establishments are concerned with protecting secrets against espionage and physical attack.
- A company headquartered in the mountains may need to anticipate avalanches more than a company located next to the beach.

Each organisation's needs will be different. Figure 7.6 (page 320) gives a list of possible measures that could reduce instances of identified risks.

Invest in outside consultants to inspect your organisation, test your security and examine your equipment and procedures to identify potential risks or weaknesses. Where possible, take preventative measures to reduce the chance of identified risks actually occurring. Buy a fireproof safe and install fire alarms; in flood-prone areas, move servers onto high racks. Train staff not to give out passwords to smooth-talking strangers on the phone. Improve malware defences. Invest in a continuous data backup scheme and install network monitoring software to detect intrusions. Every dollar spent on hazard reduction saves four times that in recovery costs.

Relying on a single solution in an emergency may turn out to be short sighted. A good DRP will make **contingency plans** if preferred procedures happen to be unavailable. For example, if staff cannot reach the emergency meeting point, they should proceed to a secondary location. If the general manager cannot be located, there should be a hiearchy established so that another person can be contacted in their stead.

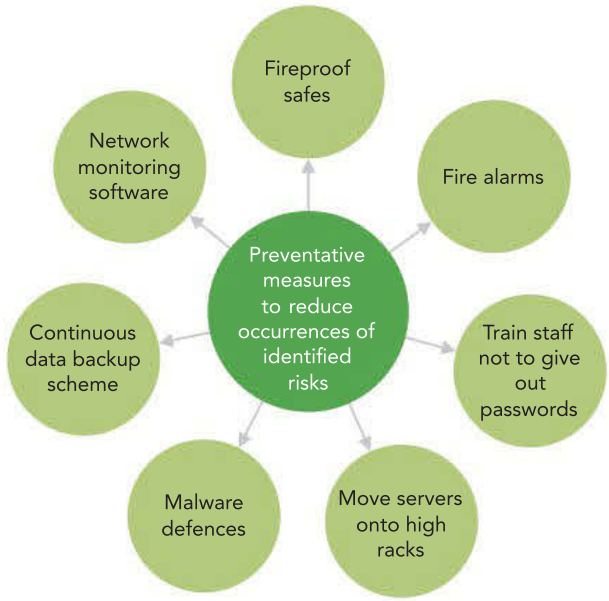


FIGURE 7.6 Reducing occurrences of identified risks

Allocating emergency responsibilities

Responsibilities during a disaster should be allocated to specific individuals, with secondary people assigned if the primary person is absent at the time of the disaster. A good DRP will state very clearly *who* is responsible for doing *what*, *where* and *when*.

Allocated tasks include:

- ongoing preventative measures, such as backing up; annually recharging and testing fire extinguishers; changing smoke alarm batteries; and updating the DRP, emergency kit, and emergency signage
- assigning roles during emergencies, such as checking that the toilets have been evacuated, shutting down servers, retrieving DRP documentation and performing head counts
- allocating jobs during recovery, such as setting up the secondary workplace, contacting the insurance company, informing clients and partners of the disaster, acquiring replacement equipment, restoring backups and testing that recovery is complete.

A large organisation would assign a disaster planning manager.

THINK ABOUT DATA ANALYTICS 7.8

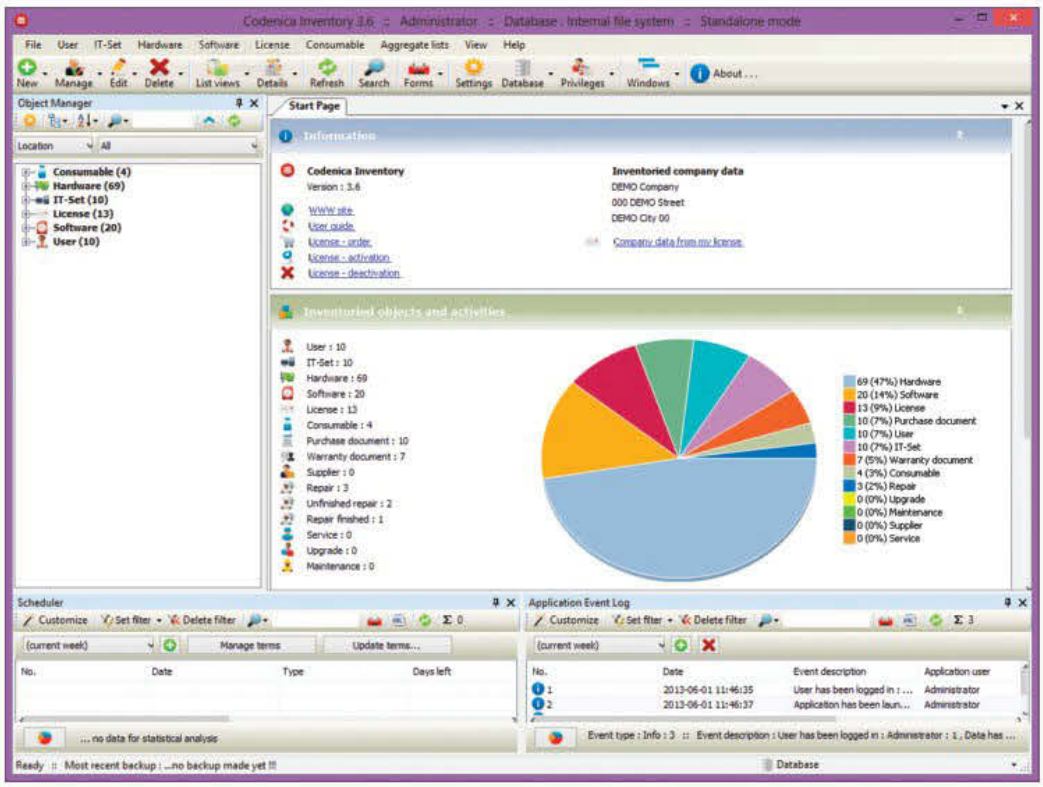
The owner of a florist, Bill, advises his brother, Ben, that he should consider their DRP confidential and not to give it out to anyone else. Why?

- Their DRP states, 'In the event of fire someone should call the fire brigade.' Bill thinks that Ben could have written this better. Reword this statement.
- What events should automatically trigger the review of a DRP, and why?

Inventory

It would be inconvenient and troublesome if you could not restore your backups because you did not know what backup software you had been using or even the operating system the computers had used. Audit all of your digital systems (hardware and software), including make, model and version numbers so that you can obtain replacements quickly. Update your audits whenever equipment changes. Print the inventory and add it to your DRP emergency kit.

9780170440875



Softpedia, http://www.softpedia.com/get/System/System-Info/Codenica-Inventory.shtml

FIGURE 7.7 Inventory software makes it easier to audit your digital systems equipment.

Key information

Document key information in printed form so it can be found easily during an emergency evacuation. Even the calmest, coolest head can become panicked or go blank in a disaster, so it is vital that the information not only be available electronically or via one person. An emergency kit that contains that key information should be created and kept in an easily accessible location, near an exit that is not likely to be blocked off in case of fire. Do not bury it in the bottom of a filing cabinet in the corner of the office that is furthest from the exit. Do not save it only electronically – you may have no way of accessing it when you need it the most.

Remember:

- print it out
- keep it near the exit
- keep a backup copy off-site, just in case.

A good DRP will assume the worst and prepare for it. It should include all the data and information you need to recover and are unlikely to be able to memorise, such as insurance policy numbers, contact phone numbers and login details. Figure 7.8 (page 322) details key information that should be included in a DRP. Remember to review the information in the emergency kit regularly to check its accuracy.

Businesses often have reciprocal agreements with other branches of their organisation, or other organisations, to provide emergency workplaces or facilities to each other.

Australians living in bushfire-prone areas know the value of a survival plan and a pre-packed kit containing emergency survival items that can be grabbed during an escape. When a fire front approaches at the speed of a car, there is no time to wonder about what to pack.

It is unwise to have only one person who knows critical passwords. If the person is injured or killed in a disaster, an organisation might find itself locked out of its own systems. Keep copies of master passwords safely stored off-site.

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

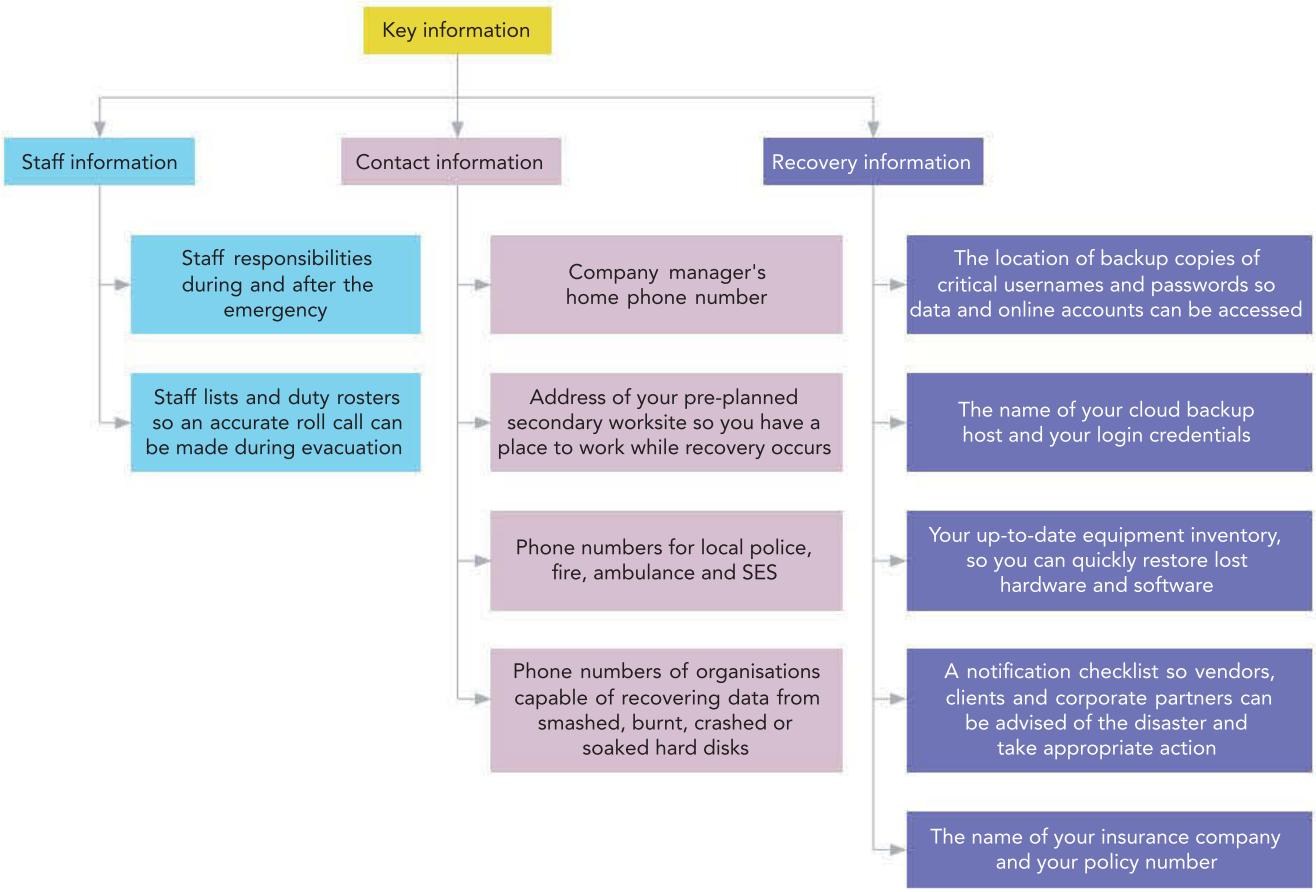


FIGURE 7.8 Key information to include in a DRP

Backup scheme

The core of a DRP must describe an effective data backup scheme: if an organisation's data cannot be recovered, little else matters. The plan should specify *how* and *when* backups are done, and by *whom*.

The most common backup scheme involves daily incremental backups and weekly full backups, as described in Table 7.6 (page 323). Other types of backups include differential backups and partial backups.

Even minor data loss can be a crisis for modern organisations. An increasingly common solution is **continuous data protection (CDP)**, which saves all versions of all data immediately when it is created or changed, and saves it off-site using the **cloud** or a remote network server. Online CDP providers you may have used personally during the course of your studies include Dropbox and Google Drive.

In the example of backup behaviours in Table 7.7 (page 323), to restore the file system, you would firstly load Monday's full backup. Next, you would either add all of the small incremental backups or add the single most recent larger differential backup.

Backups are usually scheduled to run at night because files that are open and in use often cannot be copied. Nightly backups also prevent networks being slowed down by the backup's heavy disk activity.

9780170440875

TABLE 7.6 Common types of backups

Frequency/ type	Description
Daily/ incremental	• A backup of new or changed data since the last full or incremental backup that is quick and consumes less storage space than a full backup • By itself, it will not restore a computer to its previous working state
Weekly/full	• A backup of all files in the entire system, including the operating system • Basic building block of restoring a computer to its previous working state • After data loss, the most recent full backup is restored, followed by all incremental backups that have been made since then
Differential	• Similar to an incremental backup, but saves changes since the last full backup rather than since the last incremental backup
Partial	• Only backs up a portion of the file system, such as a folder or individual files within a file system, such as a single hard disk or multiple-disk array • An incremental backup can be full incremental (saving new or changed files throughout the entire file system) or a partial incremental (saving new or changed files within a limited part of the file system)

TABLE 7.7 Comparing incremental and differential backup behaviours (document versions are shown in parentheses)

Monday			
File system	**Full backup**	**Incremental**	**Differential**
Operating system	Operating system	N/A	N/A
Doc1 (v1)	Doc1 (v1)		
Doc2 (v1)	Doc1 (v1)		
Tuesday			
File system	**Full backup**	**Incremental**	**Differential**
Operating system	N/A		
Doc1 (v2)		Doc1(v2)	Doc1(v2)
Doc2 (v1)			
Doc3 (v1)		Doc3 (v1)	Doc3 (v1)
Wednesday			
File system	**Full backup**	**Incremental**	**Differential**
Operating system	N/A		
Doc1 (v2)			Doc1 (v2)
Doc2 (v2)		Doc2 (v2)	Doc2 (v2)
Doc3 (v1)			Doc3 (v1)
Doc4 (v1)		Doc4 (v1)	Doc4 (v1)

9780170440875

TABLE 7.8 Pros and cons of backup media

Medium	Strengths	Weaknesses
Tape	• Common and popular in the business world	• Wears out with use and degrades over time • Limited capacity • Slow to save/restore data • Tapes and drives can be expensive • Can be erased by a strong magnetic field • Declining in favour of network or cloud backups
External hard disk drive	• Fast, high-capacity, cheap, portable, easily available and easy to use • Can be automated • Data never leaves your control	• Will fail over time • Can be damaged by mishandling • Can be erased by a strong magnetic field
CD/DVD	• Relatively cheap to implement • Read-only data cannot be changed later (for example, to cover up embezzlement or fraud) • Very portable	• Easily damaged and degrades over time • Limited capacity: 740 MB for CD, 4.7 GB for DVD, 25 GB for Blu-ray • Must be written manually • Read-only disks are not reusable
USB flash drive	• Small, very portable and cheap • Easily available and convenient	• With a limited rewrite capacity, they can fail without warning • USB 2.0 devices can be slow • Easily lost
Online/cloud	• Backups occur automatically, immediately and continuously • Relatively cheap • Off-site storage • Large (sometimes unlimited) capacities are available • Every version of a document over time can be recovered	• Slow to upload large files • Valuable data is entrusted to a third party • Will not work without an internet connection



Shutterstock.com/KjetilKolbjornsrud

FIGURE 7.9 A quarter-inch cassette (QIC) backup tape

Backup media

Several backup media (methods of storage) are available, each with its own strengths and weaknesses. The most popular backup media are described in Table 7.8.

Backup media should be stored off-site. It is senseless to store data backups in the same location as the originals because if a disaster destroys the master copy, it will also destroy the backup copy. This is why backing up to a LAN is not wise.

In case the backup manager is absent and someone else has to take over, the backup procedure should be documented clearly so it can be replicated by another person easily. Leaving it to individuals to figure out their own backup procedures is a dangerous practice. If the sole backup expert suddenly leaves, an organisation might find itself wondering how to make and restore backups.

9780170440875

Backups are often compressed to achieve maximum storage in a given space. They are also often encrypted so they are useless to a thief.

Automated backup software can also carry out **data deduplication** so unnecessary copies of files are not wastefully saved multiple times.

It is far more convenient for a team to backup and protect one storage location than it is to store files across the hard disks of many separate computers. This is a good argument for creating a local area network (LAN) with central document storage on the server or a network attached storage (NAS) device (Figure 7.10). Employees would be expected to save all of their work to this central networked location rather than to their local hard disks. Not only are an organisation's documents easier to backup, but they are able to be easily accessed and shared by anyone on the LAN.



Getty Images / Joby Sessions / N-Photo Magazine

FIGURE 7.10 Network attached storage (NAS) device

Testing the backup scheme

If you do not regularly test your backups, you are not really protecting your data and information. You must test your backup scheme so that you can be sure your data can be restored. Assuming that your backups will be successful without confirming with a test may be catastrophic when the day comes that you actually need to use them. You may find that your backup scheme is copying the wrong network folders, the backup media has deteriorated or the backup media is not large enough.

Some companies may have never tested a data restore previously. Backups should be tested to ensure they are still saving the right data and can be retrieved in case of emergency.

Backup schemes can go out of date, and this could be unnoticed. New servers may be installed, but not added to the backup schedule. New software or operating system changes may suddenly cause the backup software to misbehave without being detected. Storage requirements may grow to the point that the current backup media are no longer big enough, or the midnight backup is still not finished by start of business the next day, because the backup is now too large and slow. Test your backup scheme at least once a year to detect such problems.

Not even an on-site fireproof safe is acceptable for storing data backups. No safe is completely fireproof or theft-proof.

Warning: Data that has been compressed and/or encrypted requires decompression and decryption before it can be restored. Without the right software, the backup media will be useless. In the DRP, be sure to document the software (and decryption key, if necessary) that is required to restore data from the backup media.

According to the Symantec Disaster Recovery Research 2010 report, 40% of backup recovery tests fail.

TABLE 7.9 Methods for testing a backup scheme

Testing method	Explanation
Create a few typical (dummy) documents and delete them a few days later.	Can the documents be recovered from backup? Try recovering them two weeks, a month or a year later. Are they still recoverable? The loss of some data may not be noticed for quite some time if it is not used regularly.
Recover a backup of a database, video or spreadsheet that is several years' old.	Can the document still be opened after applications have been gradually upgraded or replaced over time? While most major applications try to retain their ability to open documents created in previous versions, there often comes a time when backward compatibility cannot be maintained and old documents can only be opened by old applications.
Perform integrity tests of tapes or hard disks.	This will detect media faults that could lead to corrupt data.
Completely restore from backup to another location or to a different hardware.	Remember that in a real disaster, the original hardware may need to be replaced with newer or different equipment. Will the restored system still boot after the operating system and applications have been restored? If the destination hardware is different to the current hardware, some software (for example, SQL Server, Exchange Server, Windows) may not react well.

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan | ☐ Collect complex data sets | ☐ Analysis | ☐ Folio of alternative designs | ☐ Infographic or dynamic data visualisations | ☐ Evaluation and assessment | ☐ Finalise report or visual plan

9780170440875

Failover is the process of switching data from the primary system to a secondary system (such as a backup or recovery site).

Failback is the process of returning data to the primary system after it has been shifted to a backup during failover (either because of a disaster or scheduled maintenance period).

Do not try a full test restore on your primary digital system. Downing the system and performing the restoration will stop all network activity for hours and, if it fails, your network will be broken and data will be lost, which is exactly what the test was trying to avoid.

The 'warmer' your secondary site is, the more it will cost.

Testing method	Explanation
Conduct failover and failback tests.	Some advanced systems have built-in redundancy, where all data is mirrored (copied automatically to secondary equipment). If the primary system fails, the secondary equipment activates. In such a system, you need to conduct a failover test to verify that control smoothly passes to the secondary system, and then conduct a failback test to check that control is passed back to the primary system when the disaster has been fixed.
Use a hot, warm or cold site.	Some organisations use hot sites, warm sites or cold sites where a third-party organisation provides a workplace in the case of disaster. • A hot site replicates a system and keeps a continuous, real-time copy of all data. You could walk in to the hot site and commence work immediately as if no disaster had ever occurred. • A warm site provides a clone of a system, but you must provide your own data backups to install onto the clone system when the disaster occurs. • A cold site can be configured to act as a replacement system while disaster recovery takes place, but it has neither cloned hardware nor live data backups. If you use a secondary site, you must also test that it can provide a functional workplace.

Before starting your backup test, remember to:

- plan exactly which operations will be tested, as well as where and how they will be tested
- notify staff, in case their work will be affected
- record the result of the test
- review and improve the backup scheme in the light of results of the test.

Testing the disaster plan

Organisations must regularly practise their emergency procedures and their disaster plan. This ensures that the procedures remain appropriate. It also helps staff who might otherwise be prone to anxiety become familiar and comfortable with procedures and thus less likely to add to the chaos if a real emergency occurs. Calm and confident staff will respond efficiently and effectively, but this requires drilling and repetition of procedures.



Shutterstock.com/Goygel-Sokol Dmitry

FIGURE 7.11 An emergency evacuation route map

Emergency procedures need to be documented in a form that is clear and simple, so mistakes are not made in the heat of the moment.

Staff should be assigned to specific roles so everyone knows what they need to do and no vital steps are missed. Use rehearsal drills to practise evacuation. It is important to make sure you know how many people need to be accounted for and who is off-site or absent. Evacuating to the designated assembly point regularly also reminds staff to go to the correct place. If staff scatter during an emergency, it will be difficult to account for anyone who is missing.

Practise using firefighting equipment. It is important that all staff members know where the equipment is, as well as how it works.

Perform emergency data backups. Even the best backups may be hours out of date, so new data is lost after restoring from backups. If you have any advance warning of disaster, you may still have time to copy the latest data. Always make sure you have storage media at hand that can be used for an emergency data backup. Can you quickly find where the new data is stored on the network?

Shut systems down safely. If staff panic and systems shut down without caution, there is a risk of losing unsaved data and damaging sensitive hardware. Exercise particular care with network servers.

Practise notifying emergency services. Do staff know what information to provide, such as the nature of the emergency and the address? Consider including a mobile phone in the emergency kit.

Emergency drills should be both pre-announced and surprise events, and should be conducted regularly enough that all staff are familiar with the procedures. Feedback should be collected after drills and errors or omissions in the plan should be identified and fixed.



Shutterstock.com/Cool Vector Maker

FIGURE 7.12 Simple emergency procedure documentation

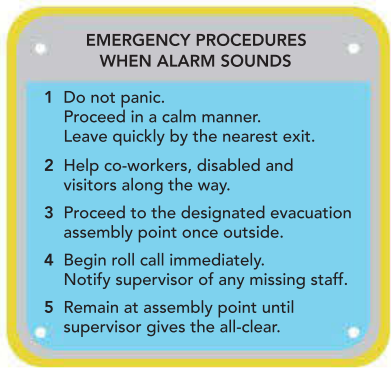


FIGURE 7.13 Text-based emergency procedures

Recovery strategies

After a disaster comes recovery. The workplace is returned to working order. Replacement equipment is acquired and installed. Data backups are restored. Complete recovery is verified according to pre-planned criteria, such as that shown in Figure 7.14.

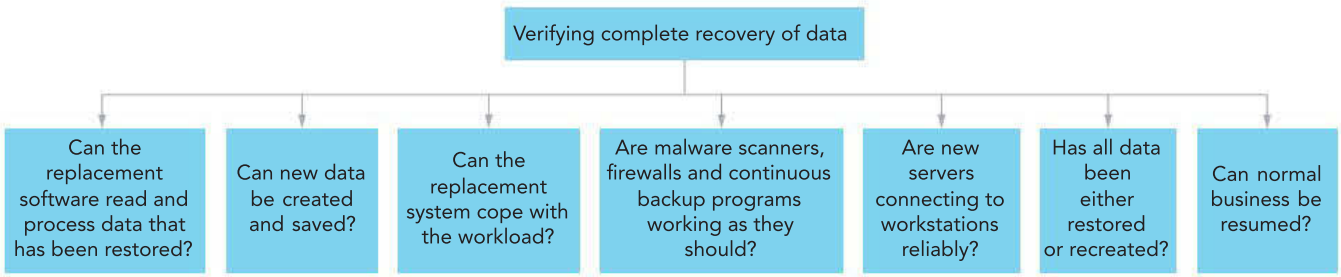


FIGURE 7.14 Criteria for verifying the complete recovery of data

SCHOOL-ASSESSED TASK TRACKER

- ☐ Project plan
- ☐ Collect complex data sets
- ☐ Analysis
- ☐ Folio of alternative designs
- ☐ Infographic or dynamic data visualisations
- ☐ Evaluation and assessment
- ☐ Finalise report or visual plan

Consequences of security failure

Organisations that fail to secure the data and information they hold, and suffer losses or breaches as a result of this, may be subject to penalties or prosecution. For example, if private personal information is lost, damaged or exposed, organisations may be prosecuted under the *Privacy Act*. If tax records are lost, organisations may be penalised or prosecuted by the Australian Taxation Office. However, the consequences of failing to protect stored and communicated data go beyond the legal and punitive.

The highly destructive Stuxnet worm was released in 2005, but not discovered as a threat until 2010. The effects of the damage over the course of those five years are unknown.

Organisations may struggle to conduct normal business because, in some cases, data loss means they are unable to pay wages to staff or pay their suppliers. Additional costs and disruptions will be caused by the need to recreate lost or damaged data, and repair or replace damaged, destroyed or stolen equipment. Labour will also be needed to address these issues. If normal business is disrupted, the organisation will also lose income.

After a security breach in 2009, Heartland Payment Systems (credit card processors) in the USA lost 50% of their market value. They accrued $139.4 million in breach-related expenses, including legal fees, forensic costs, fines and other settlement costs. It took them over a year to recover from their stock market plunge.

Data security failures may result in organisations losing their trade secrets to competitors. Depending on the level of publicity involved in the breach, the organisation may also lose public reputation as trustworthy and thus reduced customer loyalty. Even its stock market value may decline.

In 2011, Sony's Playstation Network incurred a major security failure that became one of the most famous security breaches in history, with a variety of consequences. The infographic shown as Figure 7.16 (page 329) outlines the security failure, how it unfolded and how it could have been avoided. Google 'Playstation Network Hack Infographic World' to view the infographic in detail.

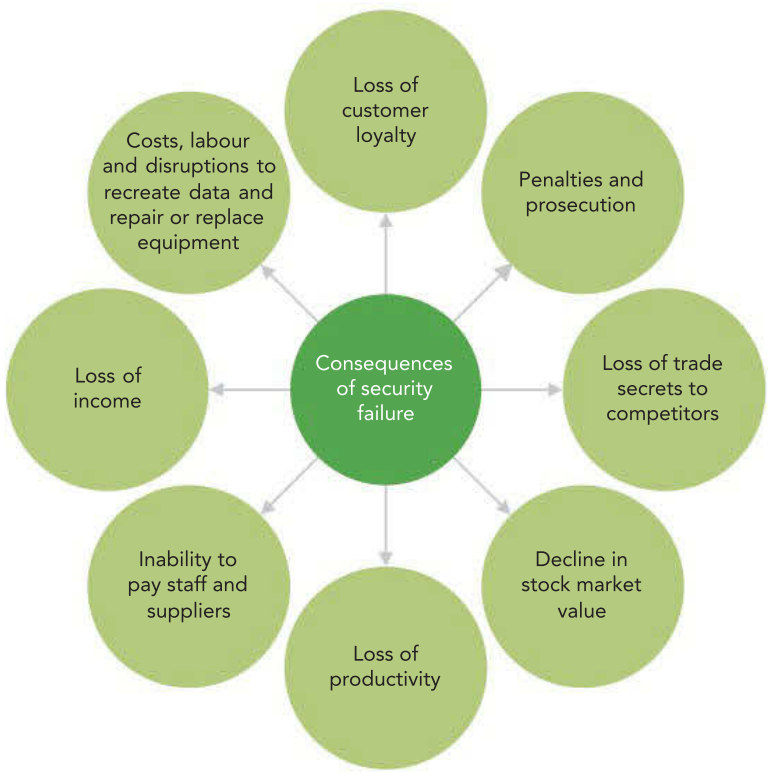
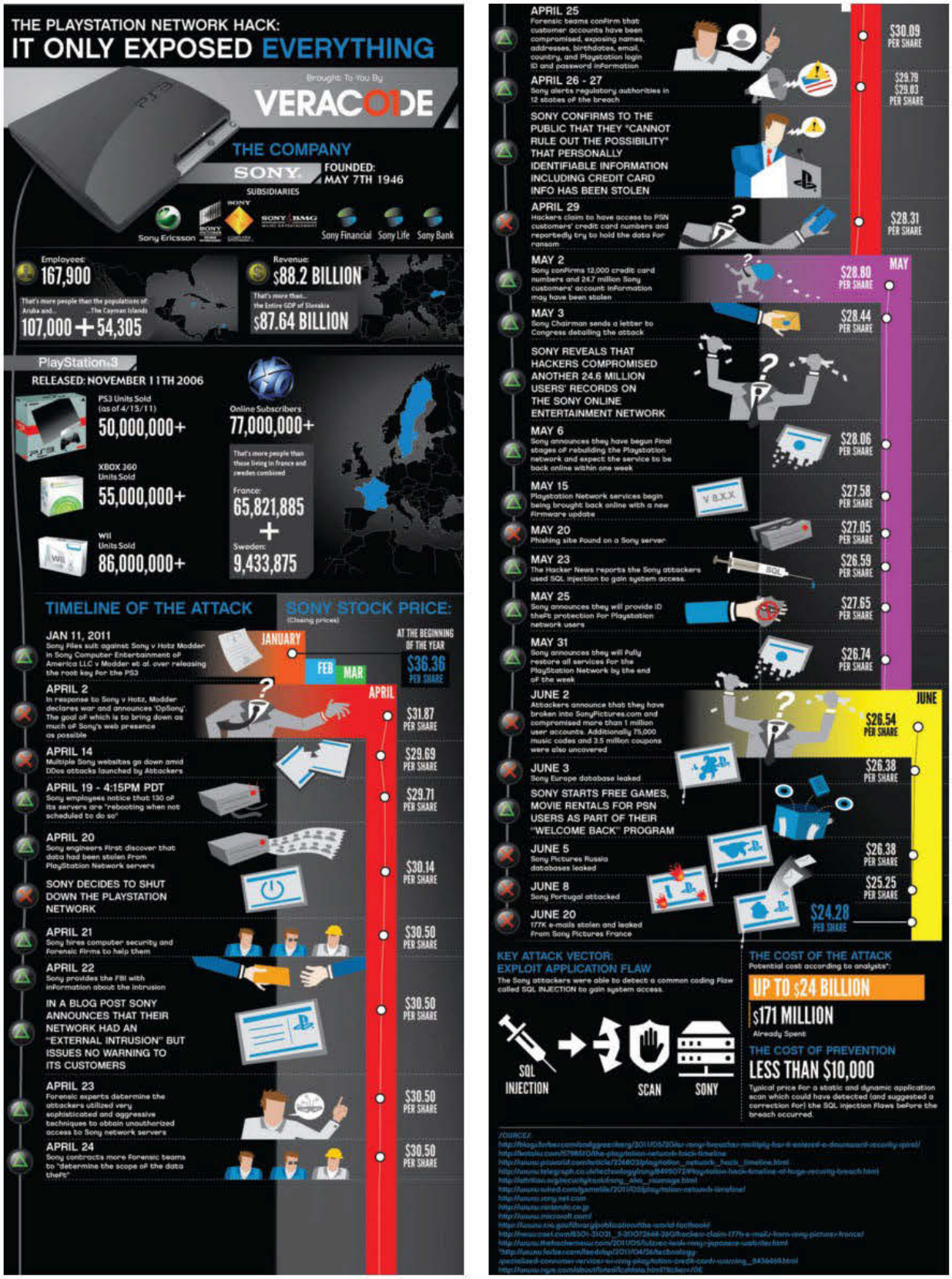


FIGURE 7.15 Consequences of data security failure to organisations



Created by Infographic World, The Visual Marketing Agency.

FIGURE 7.16 The Playstation network hack

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

CASE STUDY

Crime syndicate hacks 15 000 medical files at Cabrini Hospital, demands ransom

A cyber crime syndicate has hacked and scrambled the medical files of about 15 000 patients from a specialist cardiology unit at Cabrini Hospital and demanded a ransom.

The attack is now the subject of a joint investigation by Commonwealth security agencies.

Melbourne Heart Group, which is based at the private hospital in Malvern, has been unable to access some patient files for more than three weeks, after the malware attack crippled its server and corrupted data.

The malware used to penetrate the unit's security network is believed to be from North Korea or Russia, while the origin of the criminals behind the attack has not been revealed.

The online gang responsible for the data breach demanded a ransom be paid in cryptocurrency before a password would be provided to break the encryption.

The Age understands that a payment was made, but some of the scrambled files have not been recovered, among them patients' personal details and sensitive medical records that could be used for identity theft.

Some patients were told that their files had been lost but were not given any explanation. Others have turned up for appointments for which the hospital had no record.

The Australian Cyber Security Centre, which is part of the Australian Signals Directorate the government agency responsible for Australia's cyber warfare and information security – said it was assisting the hospital with cyber security advice.

The Australian Federal Police has also been briefed.

A Melbourne Heart Group spokeswoman said it was working with government agencies to resolve the issue.

'The protection of personal patient information is of the utmost importance ... patient privacy has not been compromised in this instance,' the spokeswoman said.

She also stressed there was no link between the encrypted data and any function relating to cardiac implantable electrical devices, such as pacemakers and defibrillators.

The spokeswoman would not say how many files had been affected or whether a ransom had been paid.

The latest hack is expected to fuel calls for the federal government to reinforce the nation's cyber defences, particularly email security.

This week, the Morrison government conceded federal parliament and major political parties' security systems had been compromised by what was believed to be a state-based cyber-attack.

Professor Matt Warren, deputy director of Deakin University's Centre for Cyber Security Research, said the data breach at Cabrini Hospital was most likely a 'ransomware' attack.

Someone, probably a staff member, using the hospital's software could have inadvertently opened a corrupted link on a phishing email allowing ransomware, a form of malware, into the hospital's system, Professor Warren said.

From there, the attackers encrypt sensitive information from hospital servers, essentially locking it away from access by medical staff.

'Then they say to the hospital "you must pay us to get your data back",' Professor Warren said.

'It's sophisticated in that you have to get the malware onto the hospital system, but once you have done that then it is relatively easy.

'Other than the cost it isn't hard to be protected from this … organisations need to update and patch their security and systems regularly because the problem you have is the hackers' capabilities are becoming more sophisticated.'

9780170440875

These types of breaches stem from the worldwide 'WannaCry' ransomware attack in May 2017.

One of the largest hit by this attack was Britain's National Health Service, where it was estimated up to 70 000 hospital devices in England and Scotland were impacted.

Non-critical emergencies and some ambulances were turned away from hospitals hit by the attack, operations were cancelled and accident and emergency centres were closed.

The healthcare sector has become a preferred target for many online criminals after the Hollywood Presbyterian Hospital in Los Angeles revealed it paid $17 000 in bitcoin to hackers who had seized control of its computer network.

And the massive hack of US health insurance giant Anthem in 2015 – when the personal information of more than 79 million Americans was exposed – further identified the sector's vulnerability to data breaches and potential for identity theft.

Houston, C. & Colangelo, A. (2019, February 20). Crime syndicate hacks 15000 medical files at Cabrini Hospital, demands ransom. *The Age*. The use of this work has been licensed by Copyright Agency except as permitted by the Copyright Act, you must not re-use this work without the permission of the copyright owner or Copyright Agency.

Evaluating information management strategies

To evaluate whether information management strategies are effective, you can apply a number of criteria. These criteria should correlate with the functional and non-functional requirements for a solution – the reasons for its development and design, as well as its ease of use. These requirements are effectively measures of success that will be monitored over time.

Monitoring may use logs and records (for example, to count numbers of errors or measure speed). It may also involve using interviews or questionnaires for opinion-based issues, such as ease of use, happiness or feeling of comfort using the information system.

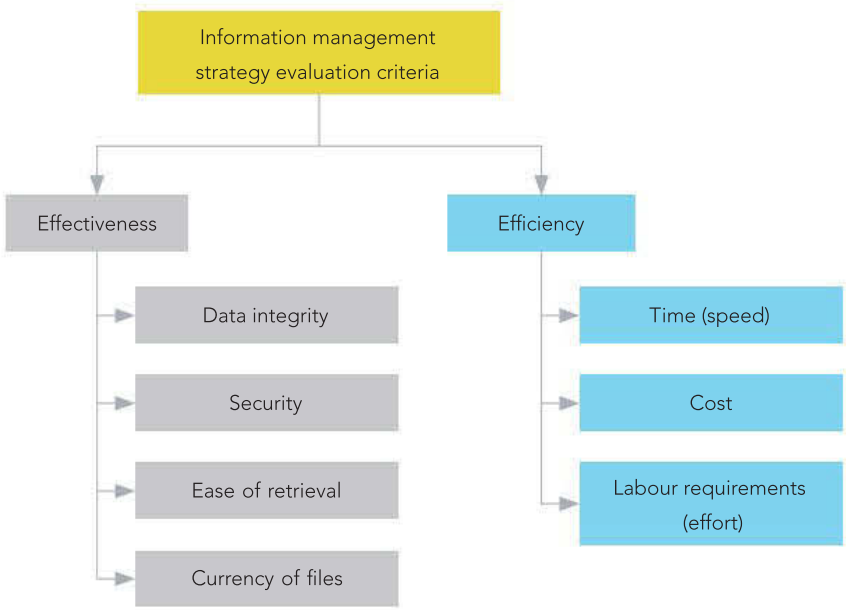


FIGURE 7.17 Information management strategy evaluation criteria

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

Evaluation does not check that a solution can cope with suddenly removing a power plug, or that it calculates the total of invoices correctly. Such issues are determined during testing before a solution is released. To reiterate: Evaluation checks that a solution is achieving the goals or objectives for which it was originally designed.

Evaluation criteria fall under effectiveness or efficiency.

Effectiveness criteria

Effectiveness criteria measure the quality of the strategy – how well it performs. The measures of effectiveness in an information management strategy are data integrity, security, ease of retrieval and currency of files. Each criterion is expanded upon in the section that follows.

Data integrity

Data integrity is data that is both consistent and accurate over the lifetime of the data. There should be few or no complaints from customers or users about inaccurate data or information about their accounts. Network logs should report few or no occasions where data storage errors have occurred. Unreasonable data should be reliably and appropriately queried or rejected.

Security

There should be no reports lodged by users about deleted data, or their data being accessed, used or revealed by unauthorised parties. Data disposal procedures should never have accidentally expose sensitive data and information. There should be no incidences of successful security breaches – from hackers, crackers or otherwise.

Ease of retrieval

Depending on the type of information system, staff surveys or customer feedback will reveal that users are happy with the ease of retrieving relevant data and information from the database. It should not be unnecessarily complicated or difficult to locate the relevant data and information using queries, searches, sorting, filtering and so on. This also relates to efficiency criteria – time (speed); that is, users should not report that it is unnecessarily time consuming to retrieve the data and information because of the complexity of retrieval processes.

Currency of files

Logs should report few or no complaints from staff or customers about data and information being out of date. This also relates to data integrity – timeliness. An example of a customer complaint for currency of files might include a university website that has class timetables listed that are more than a year old, but none for the current or upcoming semesters.

9780170440875

Efficiency criteria

Efficiency criteria relate to measurable, numeric or factual issues. Each criterion is expanded upon in the section that follows.

Time (speed)

This criterion evaluates a strategy based on its speed. Logs should report few or no complaints from staff or customers about the strategy being slow to respond, tending to lag or slow to use overall.

Cost

This criterion evaluates a strategy based on ongoing costs. If the ongoing costs of operating the strategy outweigh the benefits of using the strategy, the strategy is not efficient. Evaluation criteria can address start-up costs, but it is more likely that ongoing costs will be of more interest here. For example, if the strategy included a sophisticated type of laser printer and software within it, ongoing costs would include supplies such as toner, drum cartridges and paper, as well as power.

Labour requirements (effort)

This criterion evaluates a strategy based on the number of labour hours required to produce output. If the labour hours required to produce output are exceedingly high, they have an effect also on cost and have a negative impact on the efficiency of the strategy.

9780170440875

7 CHAPTER SUMMARY

Essential terms

cloud storing data in a remote location, using an internet connection to transfer data

code of conduct set of conventional principles and expectations that is considered binding on any member of a particular group

contingency plan a course of action designed to respond effectively to a significant event that may occur

continuous data protection (CDP) storage technology that can save real-time data changes and enable data to be restored

cookie a small file that a web server stores on a user's computer that contains data about the user (such email address and browsing preferences)

data raw, unorganised facts, figures and symbols; data can also mean ideas or concepts before they have been refined

data deduplication removing copies of (or redundant) data in a data set

data integrity ensuring that all data is trustworthy, which is achieved through accuracy, authenticity, correctness, reasonableness, relevance and timeliness

dilemma when there is a choice between two or more equally undesirable options

disaster recovery plan (DRP) a strategy developed in advance to explain how to prepare for, survive, and recover from, catastrophic data loss

ethics the principles of right and wrong, which are accepted by an individual or a social group; ethical behaviour often guides policy-makers within organisations

mission statement basis for establishing a set of common goals that will help accomplish an organisation's aims

objectives small achievable tasks undertaken to accomplish a big task

organisational goal how an organisation intends to go about achieving its mission

risk assessment analysing the probability of different risks to data and information, and the potential consequences of each of those risks

system goal the specific role of the information system in achieving the organisational goal, and ultimately the company's mission

9780170440875

Important facts

1. Information is obtained when **data** is manipulated by the computer's processor into a meaningful and useful form (also becoming output).
2. The large amount of data that is being gathered and analysed has become very valuable to organisations. Large data sets (or **big data**) assist with competition, productivity growth and innovation.
3. Data can be **compromised** through activities such as deliberate attacks of theft (disgruntled workers or criminals that will make money selling the data), loss of devices (such as accidentally leaving laptops at airports or in a taxi), neglect (not erasing data when recycling computer hardware) and not following appropriate data handling procedures and policies.
4. ***Privacy Act 1988*** was amended by the *Privacy Amendment (Enhancing Privacy Protection)* Bill in 2012. This came into effect in 2014. As part of this Act, the Australian Privacy Principles (APPs) replaced the National Privacy Principles and the Information Privacy Principles so Australia now has one set of Principles.
5. As part of the *Privacy Act*, the **Australian Privacy Principles (APPs)** were devised to set out the standards, rights and obligations for collecting, handling, holding, accessing, using, disclosing and correcting personal information. There are 13 APPs.
6. The ***Privacy and Data Protection Act 2014 (PDPA)*** was introduced by the Victorian Government. It replaced the *Information Privacy Act 2000*.
7. The other key Victorian law relating to privacy is the ***Health Records Act 2001***. This Act governs the collection and handling of confidential medical records.
8. A **disaster recovery plan (DRP)** starts with a **risk assessment** to discover the types and probabilities of threats your data could face.
9. A DRP assigns **disaster responsibilities** to individuals.
10. A DRP includes **inventories** of hardware and software that will need to be replaced.
11. A DRP contains **vital information** that will be needed after a disaster.
12. The main feature of a DRP is a **good backup** and a **data recovery scheme**.
13. **Backups** should be regular, tested, documented and stored off-site.
14. **Continuous data protection (CDP)** backs up new or changed data immediately to the cloud.
15. A DRP explains **emergency procedures** to take during emergencies, such as an evacuation plan, system shutdown and how to use firefighting equipment.
16. A **DRP** explains how to restore hardware, software and data, and how to evaluate the success of the recovery.
17. There are many serious consequences of a failure of **data security**, including loss of reputation, inability to carry on normal business and loss of trade secrets to competitors.

9780170440875

TEST YOUR KNOWLEDGE

Information laws

1 Explain why an organisation must comply with legal requirements.
2 Briefly summarise the role and scope of the three key laws affecting privacy of information.
3 Why have these laws been introduced?
4 If you believe that the privacy of your information has been breached by the Australian Taxation Office, to whom can you complain?
5 What are the penalties for breaches of the *Privacy and Data Protection Act 2014*?
6 For each of the following breaches of privacy, identify which privacy law would cover it.
   a You find that your employer has published your tax file number on the internet.
   b Medical records are found at the tip.
   c A bank refuses to give you a loan because the manager claims your credit record is poor, when it should actually be very good.
   d A consultant working for the Victorian Government passes on your VCE results to a friend without your permission.
   e A website you visit asks for personal information from you, but does not display its privacy policy.

Ethics

7 Explain how ethical requirements differ from legal requirements.
8 What are some ethical responsibilities that employers and employees have to one another?
9 What is the purpose of a code of conduct?
10 What type of ethical restrictions can be applied to accessing the internet at work?

Data security

11 Justify the need for a disaster recovery plan to a reluctant small-business owner.
12 List four major consequences if a business ignores or violates security measures, and justify your choices.
13 What are the main features of an effective backup scheme?
14 Describe five important procedures that should be used to protect data and information. Do not repeat items used in the previous question.

9780170440875

APPLY YOUR KNOWLEDGE



Charlotte runs a small business with four employees. Her business is a medical practice with a doctor, nurse and two office administrators. Her medical practice consists of four desktop computers and two laptops. Each of the desktop computers is attached to a printer and the computers are all networked wirelessly.

1. Justify why a business of Charlotte's size should have a disaster recovery plan.
2. Create a draft disaster recovery plan, including:
   - a. a risk assessment outlining the threats her data faces
   - b. a data backup scheme
   - c. evacuation procedures
   - d. data restoration procedures
   - e. a method of testing the plan.
3. Explain the possible consequences of Charlotte not taking the threats to her data seriously.
4. If Charlotte did adopt different information management practices, how could she tell whether they were better than her old practices?
5. Which information laws should Charlotte be aware of?
6. Describe the laws and the aspects that directly relate to Charlotte's medical practice and the oversight of medical records.

SCHOOL-ASSESSED TASK TRACKER

☐ Project plan	☐ Collect complex data sets	☐ Analysis	☐ Folio of alternative designs	☐ Infographic or dynamic data visualisations	☐ Evaluation and assessment	☐ Finalise report or visual plan

9780170440875

PREPARING FOR Unit 4

OUTCOME 2

Respond to a teacher-provided case study to investigate the current data and information security strategies of an organisation, examine the threats to the security of data and information, and recommend strategies to improve current practices.

To achieve this Outcome, you will draw on key knowledge and key skills outlined in Unit 4, Area of Study 2 – Cybersecurity: Data and information security. This Area of Study is covered in chapters 6 and 7 of this textbook.

Steps to follow

1. Read the teacher-provided case study.
2. Analyse and discuss the current data and information security strategies used by an organisation.
3. Propose and apply criteria to evaluate the effectiveness of current data and information security strategies.
4. Identify and evaluate threats to the security of data and information.
5. Identify and discuss possible legal and ethical consequences of ineffective data and information security strategies.
6. Recommend and justify strategies to improve current data and information security practices.

Documents required for assessment

Submit your responses to the provided case study to your teacher. This could be in the format of written responses to structured questions, a written report or a multimedia report (depending on the requirements outlined by your teacher).

Assessment

This task is marked out of 100 and is worth 10% of your study score. Your performance will be assessed using one of the following.

- A case study with structured questions
- A report in written format
- A report in multimedia format

9780170440875

Index

256-bit keys 155
3-2-1 backup strategy 217
802.11 protocol 270, 286, 288, 291

Aboriginal and Torres Strait Islander people, religion and 225
absolute referencing 86, 87, 107
acceptable-use policy, internet 317
accessibility
  checks 246
  data 312
  design 38, 92, 173
  logs 277, 291
  personal information 310
  restrictions 278, 291
  screens 234
  solutions 229–30, 249
  tests 96, 105
Accessibility Color Wheel 229
accidental threats 272–3, 291
accounting programs 169
accuracy
  data 20–2, 50
  project 255
  solutions 230, 249
  testing 105
  validation compared with 243
acknowledgements 183, 240
acquisition procedures 50, 142
active voice 227
activities performed (list) 277
adaptation strategy 182
ad blockers 231, 232
add-ins 235
Adobe Flash Player 103, 231, 233, 246
Adobe Photoshop 152, 242
adware 157, 273
age appropriate language 226
air quality data 8–9
airflow visualisations 34–5, 212
airport security 154
algorithms 187, 200, 279
alignment (text) 38, 48, 49
alphabetical sort 75, 100
alt text 92, 229, 247
American Psychological Association (APA) style guide 14–15, 44, 50, 137, 138
analogue data 151, 263
analysis 50, 120, 121
analytic coding 131, 135, 159
animations 89, 90, 92, 102, 175, 193, 209, 231, 233, 242, 245, 249
annotated diagrams 48, 77, 84, 92, 107, 188–9
annotations 90, 91, 200, 255
  project plans 251, 252
annualised failure rate (AFR), hard drive 218
anonymity, personal information 309, 312, 313, 314
anti-malware protection 157
anti-spyware protection 302
antivirus software 156, 302
appearance
  design principles 38, 48, 173–6
  spreadsheets 88
  visualisations 89
Apple Mac operating system 185, 215, 216–17, 228, 231
application servers 151, 281
archives 216, 219
Arial 40, 91
arithmetic devices 76, 84
arrows, lines and 43–4
articles 237
assessment, project plan 254–5
asymmetric digital subscriber line (ADSL) 151
asymmetric encryption 279
asymmetrical balance 176
attractiveness
  designs 92
  solutions 233–4, 249, 253
  testing 104
attribute lists 169, 181
audience inclusion 222, 249
audio clips 89, 102, 233, 245
audio files 91, 209, 238
audio editors 242
audit trails 277, 291, 304
Australian Bureau of Statistics (ABS) 5, 9, 11, 13, 58, 72, 80, 126, 138
Australian Computer Society 316
Australian Copyright Council 306
Australian Cyber Security Centre 330
Australian English 15, 138
Australian Human Rights Commission guidelines 223
Australian Privacy Principles (APPs) 308–10, 311, 314
Australian schools, data 58
Australian Tax Office (ATO) 152, 219
  penalties 328
Australia–United States Free Trade Agreement (AUSFTA) 305, 307
authentication, biometric 282
authenticity, data 10, 22, 50, 139–40
authoring software 236, 242, 255
authorised users 282, 284, 291
automated backups 325
automated cross-browser compatibility checks 103
avatars 140
A–Z sort 101

Back button 105
background checks 302
background shades 41, 42
backups 158, 217–18, 231, 280, 282–4, 291, 302–3, 319
  copyright and 305–7
  creation 290
  disasters and 320, 322–4
  locations 283
  personnel 283–4
  strategy 282, 291
  timing 282
backward compatibility, testing 325
balance, design elements 39, 176, 240
Balsamiq 152, 189
bar charts 198, 209, 212
barcode readers 148
bar graphs 26, 27, 50, 84, 191
barrier techniques 281–2, 291
base stations 271
basics emphasis 182
bathtub curve 218
batteries 272
bench tests 93, 96, 97, 101, 107, 234
bias 142, 145
bibliography 14–15, 16, 138, 139, 247
biometric security 154, 155, 282, 284, 291
birth dates 22, 146
bits 263
black swan moment 126, 159
blogs 253, 257
Bluetooth 13, 270, 271, 291
Blu-ray backup 324
bold lines 43
Boolean data 7, 17, 18, 19, 50, 136
borders 84
bots 273, 276
boundary condition data 97, 245
boxes 85
brainstorming 177–8
broadband routers 263, 264, 266
browser compatibility 103, 104, 246, 247
browsers 173, 231–2

9780170440875

brute force data attacks 276, 291
bubble charts 28, 50
budget, project completion 254
bugs 273
bullet points 198
Bureau of Meteorology (BOM) 9, 11–12, 22, 126
burglary risk 319
burned-in device 218
business intelligence (BI) 220, 256
businesses
  goals 300
  privacy and 309, 310
  support systems 299
buttons 101, 106, 175, 212, 213, 229, 247
bytes 150, 263

cable modems 151
cables 264, 267, 268
calculated fields 74, 75, 76, 78, 101, 107
calculations 68, 90
  testing 102, 246
capacity constraints 169
capital letters 67
captions 71, 72–3
cartoons 237
CAT cables 151, 269
categorisation, qualitative data 131, 133–4
CCTV 281, 284
CD ROM backup 218, 324
cells 68, 75, 85, 107, 137
  references 86
  styles 235
  validation 88
  networks 270
Census of Population and Housing 9, 11, 13
central ideas 180
central processing unit (CPU) 151, 272
central servers 264, 266
centred text 173
chain of events 180
character maps 180
character replacement 81
characters, alphanumeric 17, 18, 19, 50, 136
charts 25, 26, 50, 84, 95, 191, 200, 209, 220, 238, 239, 240
checkboxes 93, 234, 241
cherry picking 144, 159
Chicago style sheet 137
Chrome 229, 247
cipher text 279
circuit board damage 275
citations 14, 16, 44, 50, 138, 159
clarity 230, 234
  data 22, 50, 141, 145–6
  design 92
  solutions 229–30, 249
  testing 104
classroom constraints 248
clickable items 103
clients, computer professionals and 316
client/server networks 266–7, 285, 291
climate data 10, 11, 192
closed questions 7, 129, 159
cloud backups 158, 217, 280, 283, 290, 308, 322, 324, 334
  copyright and 307
  health records storage 314
codecs 103, 246
codes of conduct, computer use 315–16, 334
coding xiv, 246
cold backup sites 326
colloquialisms 225–6
colour 91
  choices 92, 240
  combinations 174, 246
  contrast 41, 101, 229
  legend coding 45
  manipulation 233-4
  scales 32–3
colour vision deficiency (CVD) 169, 174, 200
column charts 209
column graphs 27, 50
columns 85, 107
combination strategy 182
comma-separated value (CSV) files 82, 83, 86, 107, 171
Commissioner for Law Enforcement Security Act 2005 (Vic.) 311
common file formats 82–3
communication hardware 148, 151, 159
communication networks 263–71
company expansion 300
comparison strategy 182, 197
compatibility tests 96
complaint lodging, privacy 310
completeness 233
  backup restoration 325
  data 21, 50, 141, 142–5
  projects 255
  solutions 91, 249
  testing 104
component design life (CDL), hard drive 218
component tests 96
comprehensive management plans 214, 256
compressed audio formats 18
computer clock 247
computer games, copyright and 311–12
computer monitors 148, 149
computer networks 34
  codes of conduct and 315–16
computer-generated imagery (CGI) 140
concatenated primary key (PK) 61, 62–3, 65, 66
concepts, project management 115, 159, 180
conclusions, substantiation 247, 249
concurrent tasks 117, 119, 159
condescending language 225, 256
conditional formatting 32–3, 50, 100
confidentiality 300
conflict resolution, workplace ethics 317–18
conflicts of interest, computer professionals 316
connection reliability 268
consistency
  data 21, 50, 141, 146–8
  design 92, 174, 233
constraints 37, 50, 171, 186
containers, dividers and 42, 43
content filtering 280
contingencies
  disaster recovery 319
  Gantt chart 119
continuous data protection (CDP) 322, 334
contrast (visual) 38, 41, 42, 102, 174, 245
controls 127, 159, 229
conventions 50, 95, 237–9, 256
  graphic design 39
  spreadsheets 88
cookies 3, 231, 232, 317, 334
copper cables 267
copy and paste 147, 147
copy protection 304–5, 307
copyright 14, 50, 140, 170
Copyright Act 1968 (Cth) 304–7, 317
Copyright Amendment (Digital Agenda) Act 2000 (Cth) 310, 311
Copyright Amendment Act 2006 (Cth) 310
correctness, data 22, 50, 141–2, 147
cost–quality considerations 186
costs, security failures 328, 333
counter-evidence 143, 144
cracked software 140, 332
creative designs 123, 185
creative thinking tips 182–5
credit card information 278
credit reports 308, 310
criminal deception 142, 330–1
critical paths 118, 123
critical thinking focus 182
cross-border information disclosure 309
crosscut shredders 284
cross-platform compatibility 215
cultural respect 223–5
currency 88, 147
custom input devices 148
customers
  customer movements 34
  ID number 60
  loyalty loss 328

9780170440875

customer orders 61, 62–4, 65, 66
cyber security 157, 175, 314

daily backups 272, 284
dashboards 169, 172, 220, 241, 242
data 50, 114, 159, 334
  access 219–20
  accuracy 141–7
  acquisition 4, 79–80
  age 147
  analysis 124, 170, 301
  definition of 3–4
  backups 307, 318, 319, 320, 327, 334
  breaches (cloud storage) 314
  cleansing 80–2, 107, 142, 159
  collection 6, 124
  constraints 3
  content functionality 141
  corruption 274
  display 94–5
  duplication 334
  elements 60
  encryption 155
  ethics 314–17
  exports 83
  extraction 84
  falsification 139
  fields 19, 62
  formats 18, 71, 72–3, 147
  gathering 37, 171, 213
  importation 74
  inaccessibility 143
  information systems and xvi, 298–300
  integrity 20–3, 50, 60, 104, 107, 139–51, 308–9, 332, 334
  interpretation 118, 128, 168
  loss 142, 272, 302, 303, 322, 323, 334
  network management 284–9
  manipulation 234–6, 238–42
  mining 10, 301
  obsolescence 142, 143
  packets 286
  privacy 170
  quality 7, 142, 312, 313
  range 88
  redundancy 74
  relationships 59
  removal 81–2
  repositories 9–10
  requirements 169–70
  restoration 282, 318
  security 152–8, 312, 313, 332
  servers 281
  sets 10, 82–3, 84, 126, 169, 213
  sharing 264
  sources 10, 44, 50, 105
  standardisation 142
  structures 19–20, 47, 136–7
  summaries 78
  theft 153, 266, 283, 304, 311
  transfer speeds 268, 269, 270, 284
  trends 84
  types 16–18, 19, 50, 69, 70, 71, 72–4, 88, 136–7
  updates 142, 211
  validation xiv, 20–1, 23–5, 137, 172, 252, 308
databases 19, 50, 58–9, 137, 151, 169, 182
  components 67–78
  design 47
  management systems 123, 267
  queries 68, 100
  servers 267, 285
  spreadsheets and 84
  structure 59–66
  tables 46, 74, 75
data dictionaries 46, 50, 70–1, 73
data entry
  errors 88, 142, 144, 147, 147
  operators 316
data visualisation 25–36, 50, 68, 82, 89–95, 100, 169, 186, 2190–3, 198, 200, 230, 256
  effectiveness 221–34
  software 220
  solutions 248, 250
  testing 102, 103, 104, 152, 188, 209–13, 214, 233, 246–7
dates 17, 18, 19, 136, 137, 145, 215
  deduplication 325
  documents testing 325
deadlines 169, 252
deceased persons, *Health Records Act* and 312
  decision-making 84, 301, 318
decompression software 325
decorative typefaces 40
decryption 273, 279, 307, 325
dedicated servers 267
default values 236
deletion, data memory 218–19
deliberate threats 273–4, 291
denial-of-service attacks 147
dependencies 115, 116, 117, 118
  GanttProject 121, 122
dependent variables 124, 125, 127, 128
descriptive coding 131, 159
descriptive names 67
design 97, 107, 121, 248
  effectiveness 214
  ideas 176–87, 00
  infographics 198
  pre-designed elements 236
  principles 37–9, 50, 169, 171–6
  teams 254
  techniques 176–80
  testing 104
  tools 46–9, 187–9
design briefs 37, 50
  desktop computers 232, 265, 271
desktop users 281
destruction, information 319
diagrams 194
dictionaries 243, 276
differential backups 307, 322, 323
digital authenticity 22, 140
digital object identifiers (DOIs) 16, 139, 159
digital signatures 140, 142
  digital systems xvi, 147, 148–51, 159, 314
digital-to-analogue conversion 286
digitisation 130, 263
  copyright and 306–7
direct marketing 308, 309
directory structure, storage 216–17
disaster recovery plans (DRP) 217, 256, 318–27, 334
disclosure, health records 313
discrimination 223–4
disk rot 142
distributed denial of service (DDoS) attacks 156, 273
dividers, containers and 42, 43
document shredding 284, 292
documentation xv
  backups 324
  creation 215–16
  editing 242
  Gantt charts 119, 123
  project management 214, 250–5
  testing 101, 248
  verification 140
domain name system (DNS) 147, 289
dot point presentations 229
dotted lines 43
downloaded data 79, 311
doxing 153, 159
drag and drop 217
drawing tools 184, 188
drive erasure 219
Dropbox 155, 158, 322
dropdown menus 24, 101, 211, 212, 220, 239, 240, 248, 252
due diligence 315
dummy document testing 325
duty of care, workplace 315
dynamic data visualisation 169, 171, 173, 238, 239, 240–2, 249, 256
  definition 211–13
dynamic features, testing 103, 255, 246
dynamic versatile disks (DVD) drives 218, 256, 324

ebook readers 149
economic constraints 37, 169, 200
educational materials, copyright and 306

9780170440875

effectiveness 51, 101
  criteria 332
  designs 91
  solutions 249
  testing 103
efficiency
  criteria 332, 333
  data collection 6, 51
  designs 90
  efficiency testing 103
  solutions 249, 250, 253–4
electronically recorded works 304
electronic commerce 307
electronic data gathering 310
electronic mind maps 179
electronic self-defence 157, 272
electronic sensors 8–9
electronic validation xiv, 24–5, 88, 107, 243
element placement 48, 49, 91
ellipses 15, 138
emails
  employees and 316
  servers 217, 266, 285
  validation 140
embedded systems 289
emergencies 327
  health records and 312
  responsibilities 320, 321–2
employee monitoring, ethics 316–17
employers, copyright and 306–7
encryption 140, 273, 278–9, 281, 291, 305, 307, 310, 325
en dashes 16, 139
end-users 69, 70, 178, 250
English language, varieties of 225–6
enterprise computing systems 299
entities 68, 107
entity relationship diagrams (ERDs) 60, 74
entry modification 80
equipment inventory 37, 169, 320, 321, 322
error prevention 96, 172, 234
error tolerance 92, 105, 173
Ethernet 269, 288, 291
ethical issues 130, 334
  security and 314–18
euphemisms 225
evaluation 178
  criteria xiv, xv, 248, 255, 256
  information management strategies xv, 249–50, 331–3
  solutions 248–50
event-based threats 274–5, 291, 318
events, project management 118, 159
executable files 151, 156
existence check 24, 70, 88, 107, 243, 256
experimentation 185
external data 79, 151, 158, 245, 324
extrapolation 29

face recognition 282
facts, opinion and 3, 125
failover/failback tests 326
failure rates, hard drives 218
fair test 127, 159
fair use (copyright) 14, 310, 307
fake accounts 140
false negatives and false positives 157
feedback 101, 231, 249, 250
fibre-optic cable (FOC) 151, 267, 269, 287, 291
fields 60, 68, 69, 70, 71, 72–4, 75, 80, 81, 107
file allocation tables (FAT) 219
files
  backups 157–8
  deletions 157
  disposal 218–19
  management strategies 107, 214–20, 332
  organisation media 219–21
  recovery 219, 322
  servers 264, 266, 283, 284, 285, 289
file transfer protocol (FTP) 289
FileMaker 19, 24, 137, 151, 243
filenames 123, 215, 216, 272
filters 74, 75, 83, 101, 107
financial goals 299
Find and Replace 81
findings statements 240
fingerprint scanning 140, 142, 154, 155, 282
Firefox 228, 247
fireproof safes 308, 319, 320, 325
fire threats 283, 320, 326, 327
firewalls 156, 280, 281, 291, 304, 307, 327
first normal form (FNF) 60, 61–2, 107
flags 224
flat file database 58, 60, 73, 107
flatbed scanners 148
flexibility
  design 38, 172
  functions 93
floating point data 17, 19, 136
floppy drives 218
flow visualisations 34–5, 51, 193–4, 200, 240
flowcharts 89, 188
fonts 40, 91, 169, 189, 253
  see also typefaces
footers 84
forbidden sites, employers and 317
foreign keys (FK) 65, 69, 107
form (appearance) data 77, 107, 141
Form Tools 241, 242
formal tests 97
  GanttProject 121
formats 39, 51, 78, 95, 137, 200, 237–8, 256
  spreadsheets 88
  visualisations 194
formulas 84, 85, 100, 101, 188
formulation, research question 125–6
frames, Ethernet 288
fraud 142
freeform answers 130
freeform shapes 42
frequency distribution tables 128, 159, 168, 200
full backups 283, 284, 291, 307, 322, 323
full text only 229
functionality 85
  testing 104
  design 38, 92, 172–3
  infographics 198
  reports 78
  requirements 169, 171, 176, 200
  testing 100, 246
future date triggering 98–9
fuzzy logic 147, 159

games 181, 238
Gantt charts 114–15, 116, 117, 118, 119, 220–3, 152, 159, 180, 213, 239, 251, 252, 253, 255
Gapminder 12, 30
gender inclusiveness 222–3
generations 2G–5G, cellular transmission 271
Geographic Information System (GIS)
  data 9, 10, 13, 31–3, 51, 82–3, 235
  infographics 196, 200
geometric shapes 42
geospatial visualisations 33, 82, 191, 200
GIF files 233, 237
gigabytes 150
Gimp 152, 242
Google 183, 231
Google Charts 25, 94, 152
Google Docs 94
Google Drive 158, 251, 322
Google Maps 196
Google Sheets 19, 25, 82, 94
government agencies 11–12
  data privacy 307, 309
grading criteria rubric 136
grammar checkers 240
graphical user interface (GUI) 93, 148, 234
graphics 175, 198, 209, 211, 240
  browsers and 231
  organisers 179, 180
  solutions 214, 232, 238–2
  tablet and stylus 148
graphics card 272
graphics editors 188

9780170440875

graphs 25, 41, 51, 84, 102, 128, 159, 168, 191, 194, 200, 237, 239, 245
greying out 93
grids 30, 240

hackers 153, 154, 268, 280, 305, 306, 332
  medical files and 330–1
  passwords and 276
hand-held computers 271
handwritten calculations 101, 248
hanging indent paragraph style 139
hard disk drives (HDD) 150, 218, 256, 220, 283
  damage 272
  failure 217–18
hardware 148, 266
  constraints 232
  data management 284–7
  threats to 272, 273, 291
  security controls 281–4, 291
  software and 264, 291
Harvard style sheet 137
headings 84, 87, 91, 174, 175, 235, 245, 253
Health Privacy Principles (HPPs) 308, 312–13
Health Records Act 2001 (Vic.) 301, 304, 312–14
Health Services Commissioner (Vic.) 313
heat maps 94, 95
Help files 93
hierarchical infographics 197, 200
hierarchical structures 34, 42, 47, 51, 116
  GanttProject 120
histograms 25, 26–7, 51, 191
historical data 130, 193
hoax emails 274
home networks 266
hot backup sites 326
HTML format 79, 103, 235, 246, 281
hubs 265
humour 225
hyperlinks 90, 102, 238, 239, 240, 245
hyphenated numbers 80

iCloud 158, 242
identification codes 148
identifiers, government-related 309, 313
identity theft 153, 330
identity uncertainty 140
ID field 101
illustrations, age level and 227
images 18, 51, 38, 89, 136, 152, 209, 211, 233, 238, 240, 242
  discrimination and 225
  formats 237
  loading 232
  location 189
implantable electronic devices 330
importation, downloaded data 79
inauthentic data 139
incomplete data 144–5
incremental backups 254, 283, 291, 307, 322, 323
independent variables 124, 125, 127, 128
indexes 238, 247
individuals, privacy and 308
industrial nodes 83
Infogram 25, 95
infographics 36, 43, 51, 95, 170, 171, 173, 175, 186, 188, 189, 194–7, 211, 214, 221–34, 236, 238, 239–40, 245, 250
informal testing 96, 97
Information Commission 308
Information Privacy Act 2000 (Vic) 311
Information Privacy Principles (IPPs) (Victoria) 311–12
information systems 3, 4, 51, 130, 159, 182
  data and 298–300, 313
  infrastructure breakdown 318–27
  management strategies 301–2
information theft 319
informational infographics 195, 200
infrared networks 270
inkjet printers 149
input–process–output (IPO) charts xiv 46, 84, 90, 107, 187–8
inputs 16, 24, 46–7, 51, 58, 68, 69, 70, 79–80, 107, 118, 148–50, 159, 187, 243, 308
insider threats 303–4
Inspiration 152, 179, 180, 188
installation limitations, wired networks 96, 268
insurance, disasters and 320, 321, 322
integers 17, 19, 51, 136
integration tests 96
integrity, data 6
intellectual property (IP) 14, 51, 137, 304
interactive charts 89
interactive data visualisation 169
interactive online infographics 36
interactive time visualisations 30
interactivity xvi, 47, 68, 89, 92, 172, 240, 241
  social media 10
  testing 103
  web pages 229
interfaces 92, 188, 233
internal hard drives 158
international standard (ISO) format 145, 147
internet
  access 286
  addresses 289
  backups 324
  cloud 266
  connections 264, 265, 266, 267, 290
  domains 147
  functions 211
  security 307
Internet of Things 271, 292
internet protocol (IP) addresses 3, 147, 277, 288–9
internet use, ethics 317
internet users, ages 27
interpolated data 29, 144
interpretation, data 4, 10, 136
intersex audiences 222, 223
interviews 7–8, 51, 79, 118, 132, 136, 147, 159, 249, 315
invalid data 97, 245
inventory audits 320–1
iPad 185, 247
iris pattern identification 154, 282
Items field 63, 64, 66

jargon 92, 107
JavaScript 151, 231, 232, 246, 247
Jobs, Steve 185
JPG files 233, 237
justified text 173–4

Kaspersky Secure Password Check 154
kerning 92, 234
keyboard 148, 236
key fields 68
key information, disasters and 321–2
Keylogger 156, 273
Keynote 188, 235, 239, 242
keypads 284
keystrokes 273
knowledge-management systems 299

labels 84, 173
landscape orientation 172
language, discrimination and 224–5
laptop computers 155, 265, 271
laser printers 149
lateral thinking 181
layers 240
layout diagrams 48, 68, 84, 107
leading 92, 234
left-aligned text 173
legal constraints 37, 170
legal tensions, resolution 317–18
legend (documents) 45
legislation
  data and information 304–14
  information systems 301
LibreOffice 137, 151, 240
lifetime, hard drives 218
light colours 176
light-emitting diodes (LEDs) 269
Likert scales 7, 130, 135–6, 159
Lindt Cocoa Foundation 299
line charts 26, 198, 210
line graphs 27–8, 51, 191
line of best fit 28–9, 51

9780170440875

line-of-sight access 271
lines, arrows and 43–4
linked tables 82
links 101, 142, 173
liquid crystal display (LCD) 149
literary works, software and 307
live streaming 94
loading times 102, 246, 247
local area networks (LANs) 264, 265, 266, 267, 286, 288, 289, 292
  backups 324, 325
  firewalls and 280
location matching 147
log files 317
logins 246, 277, 321
  limitations 276
  passwords 154, 307
log out 155
log recovery 332, 333
Lucidchart 152, 179

Mac computer 187
MacOS 188, 189, 215, 216, 217, 219, 228, 233, 241, 242, 244, 247, 287
macros 241, 246
maintainability, project 255
maintenance, wired networks 268
malware 147, 156–7, 273, 274, 280, 281, 292, 304, 307, 308, 319, 320, 327, 330
mandatory conventions 238, 256
manual entry data 58, 79, 102, 103, 108, 310
manual validation 23–4, 88, 108, 243, 246
many-to-many data relationship 59, 108
many-to-one data relationship 59, 108
maps 25, 33, 51, 191–2, 196, 200
Mathematica 209, 220, 231, 234, 238, 239
mathematical functions, spreadsheets 84
matrix visualisations 30–1, 51, 194, 200
mean time between failures (MTBF), hard drives 218, 256
media
  constraints 233
  plug-ins and 89–90
  testing 102, 245
media-rich files 233
megabyte 150
menus 48, 49, 93, 173, 236
message clarity 39, 106
Microsoft Access 19, 24, 58, 60, 74, 80, 81, 137, 151, 235, 236
  query design 76
Microsoft Excel 19, 24, 25, 26, 29, 74, 79, 82, 86, 88, 94, 147, 151, 209, 220, 231, 234, 235, 236, 238, 239, 241–2, 251, 252
  Open XML format 82
Microsoft Office 29, 152, 158, 235, 236
Microsoft PowerPoint 188, 231, 235, 239, 241, 242
Microsoft Windows 188, 215–17, 228, 241, 242, 287, 325
Microsoft Word 14, 16, 24, 26, 138–9, 151, 152, 188, 226, 235, 240, 242, 243, 249, 251
microwave networks 270
milestones 115, 118, 119, 123, 252, 254
mind mapping 34, 128, 152, 159, 168, 178–9, 200
misinterpreted data 22, 80, 145, 146, 147
mission statements 298, 300, 334
mobile networks (cellular networks) 270, 292
mobile phones 154, 265, 270, 271
mobility limitations, wired networks 268
mock-ups 48, 51, 84, 108, 188–9, 239
modems 151, 263, 286–7, 292
moral rights 14
motion charts 94
mouse 148
mouseover events 103
moving images 175
MP3 files 18, 233, 306
multi-choice dropdowns 130
multimedia authoring tools 89, 242, 255
multiple backups 307
multiple-choice questions 135
multiple fields 74
multiple network devices 151
multiple tables 72
music files 75, 238, 307
myGov.au 154
My Health Record 314

naming conventions 67–70, 108
narratives 194
National Aeronautics and Space Administration (NASA) (US) 10, 12, 178
National Broadband Network (NBN) 192, 269, 287
navigation 92, 105, 173, 240
network architecture 266, 292
network attached storage (NAS) 150, 289–90, 292, 325
network connection speeds 169
network diagrams 25
network hardware 151
network interface card 285, 292
network monitoring 320
network nodes 289
network operating system (NOS) 287, 292
networks 263–71, 284–9, 292
network servers 327
network users 266, 267
network visualisations 192, 200
neutral language 225
nightly backups 322
nodes 265, 267, 292
non-digital media 151
non-documented testing 97
non-functional requirements 169, 171, 176, 200
non-government organisations, privacy and 307
non-judgemental brainstorming 177
non-key fields 62
non-numerical data 34, 130
non-primary key field 62
non-technical constraints 169–70, 200
non-textual information 102, 245
normal business, resumption 327, 328
normalisation 74, 82, 108
no spaces 67
notebook computers 263, 264, 271
Notepad 82, 242
notes and comments 84
null fields 220, 256
Numbers 19, 25, 82
numeric data 17, 18, 19, 26, 51, 74, 75, 100, 130, 136, 191

objective data 48, 49, 125, 129, 130, 224–6, 249, 256
objectives 301–2, 334
observations 8, 51, 79, 124, 184–5
Official Secrets Act 143
off-site backups 217, 283, 290, 304, 321, 324
one-to-many data relationship 59, 108
one-to-one data relationship 108
online
  backups 324
  Census (2006) 11
  chat 264
  surveys 7
on-screen information 188
on-site storage 290
open questions 7
openness, information disposal 312, 313
open-source software 152
operating systems 151, 157
opinion, fact and 125
opportunity maximisation 301–2
optimism focus 182
optional conventions 238, 256
options evaluation 318
Optus 151, 266
Orders items 61, 62–4, 65, 66
organic light-emitting diode (OLED) 149
organisational goals 301–2, 334
organisations, data gathering and 3, 4
outdated data backups 147, 324, 325
outputs 46–7, 148–50, 159, 187, 298
overlooked data 145
overwritten files 219

packet switching 288–9
Pages 240, 242
paper records, damage to 143
paper surveys 7
paraphrasing 137, 159

9780170440875

partial backups 283, 322, 323
passive voice 226–7
passwords 154, 155, 273, 275–6, 292, 304, 306, 319, 320, 321
patronising language 227
patterns 42
PDF documents 231
peer-to-peer network 266–7, 292
penalties
  copyright infringement 307
  data losses 328
  health services records offences 313
  privacy offences 311
peripheral devices 151
permissions 278, 307
personal information management 308, 309–11
  privacy and 274, 310–11
personal pronouns, gender considerations 223
personal use, copyright 306
'Pesticide Planet' 40, 43, 44, 45
phishing 274, 292, 330
photographs 235, 238, 315
PhotoShop 253
physical security 153, 281–2, 302
pictographs 210
pie charts 26, 30, 51, 84, 133–4, 180, 191, 209, 237
pixelation 225
placeholding text 48
plagiarism 14, 51, 140
plain text files 82
plug-ins 89–90, 102, 103, 231, 245, 246, 256
PNG files 233, 237
policy development 130
POOCH (Problem, Options, Outcomes, Choice) tools 180
pop-ups 231, 232
portable hard drives 283, 284
portrait orientation 172
ports 151
power loss 272, 292
power surges 275, 292
predecessor tasks 117, 159
preferred conventions 238, 256
prefixes 67
preliminary research 126
preparation 184–5
presentation software 238–40
preservation copies (library books) 305
Pretty Good Privacy (PGP) 156, 160
preventative measures, disasters 319, 320
primary data 4–6, 10, 51, 79, 116, 128–9, 160, 170
primary key (PK) 60, 61, 62, 65, 69, 74, 77, 108
printed output 148, 182, 188, 255, 267, 285
printers 148, 271
  networks and 263, 264–5
Privacy Act 1988 (Cth) 152, 301, 304, 304, 307–11, 314, 317, 328
Privacy Amendment (Enhancing Privacy Protection) Bill 2012 (Cth) 308
Privacy and Data Protection Act 2014 (Vic.) (PDPA) 301, 304, 309, 311–12
Privacy and Data Protection Commissioner (Victoria) 311
privacy
  data and 126
  health services and 309
  legislation 304
  organisations 311
private keys 279
problem analysis xiii, 37
problem solving xvi, 37, 186, 248, 317
problem-solving methodology (PSM) xiii–xv, 4, 37, 46, 51, 90, 97, 103, 166, 178
processes, project management 46–7, 116–23, 160, 188
process infographics 196, 200
productivity
  losses 328
  solutions 249
profitability 249, 250, 300, 301
programming languages 151
programs 238
project management 114–23, 248
  documentation 250–5
  software 115
  strategy 214
projectors 264
proof, data 126–7
proofreading 225, 243
protocols 266, 280–1, 286, 288–9, 292
proxy servers 267, 285
pseudocode 188
pseudonymity 308, 309
public keys 156, 160, 279
punctuation 225
Python 151, 234, 246

qualitative data 35, 118, 129–30
quantitative data 118, 129–30, 160
quarantining 281
queries 74, 108
  data patterns and 83
  design 75–7
  testing 99
questionnaires 115, 119, 124, 130, 132, 134, 144, 147, 249, 315
quotations 14, 15, 138
QWERTY keyboard 148

radio
  data 270, 271
  surveillance 305
radio buttons 93, 213, 220, 234, 239, 241
random access memory (RAM) 149, 150, 169, 272
random data loss 145
range checks 24, 70108, 243, 256
ranking ladders 180
ransomware 157, 273, 330–1
Raspberry Pi 182, 183
read only files 324
readability
  design solutions 91
  project 255, 256
  solutions 230, 249
  statistics 226
  testing 102, 104, 245, 247, 248
Ready Set Go! gym 132–4
reasonableness, data 23, 51
receiving devices 263
records 69, 75, 108
recovery strategies, verification 327
redundancies 108
redundant array of independent disks (RAID) 218, 290
references 14, 16, 137–9, 213
referential integrity 60, 147
refutability, data 126–7
rehearsal drills, disaster plan 327
rekeying, avoidance of 146, 147
relational database management system (RDBMS) 17, 58–9, 60, 68
relational databases 72, 74, 108, 147
relative referencing 86, 87, 108
relevance
  data 23, 51
  solutions 230, 249, 256
  sources 141
  testing 105
religion, sensitivity and 224
remote location data 290, 308
repetition, design elements 174
reports design 77–8, 108
requestioning 147
required fields 69, 70
research 183
  articles 35
  Australian Human Rights Commission guidelines 223
  data sets 118
  findings 238
  functional and non-functional requirements 176
  potential research question and data sources 129
  qualitative data coding 131–4
  repetition 175
  significance 125
research questions 116, 124–8, 129, 2170, 171, 213, 240
resources allocation 115, 118–19
results 100, 168
retinal pattern identification 154

9780170440875

retrieval ease 332
ribbon customisation 241
right-justification 174
risk assessment 185, 304, 319–20, 334
robustness (design) 38, 172
rootkit 273
rose (coxcomb) chart 209, 210
rough sketches 48
routers 265, 286, 287, 292
rows 85, 108

sabotage 308
sample pages 247
sample size 22, 51, 142
sample test data 97, 99
sans-serif typefaces 40, 91, 237, 256
satellite networks 270
scalability 186
scanners 264
scatter graphs 28–9, 31, 51, 209, 212
schedule overruns 255
schools
  computer code of conduct 316
  goals 298
  numbers 72, 87
  vandalism and 319
scientific approach 130
screen recording software 249
screenshots 101, 102, 246, 248, 249, 255, 305
scrolling 103, 173, 241
SD cards 150, 218
searching 74, 108
second normal form (2NF) 60, 62–4, 108
secondary data 5–6, 10, 51, 116, 128–9, 142, 160, 170, 326
secondary storage 149, 150
secondary worksite plans 320, 321, 322
secret keys 279
secret question information 154
secure sockets layer (SSL) 140, 156, 281
security
  constraints 37
  certificates 140
  controls 275–84, 292
  ethics and 314–18
  failure consequences 328–31
  guards 284
  personal information 309
  protocols 280–1, 288, 292
  settings 273
  systems 304–8, 332
  wired networks 268
selectable options 173
self-replicating software 274
sending devices 263
sensitive information, privacy and 312, 314
sensors 8–9, 13, 52
sentence length 230
sequencing 117, 120, 180
serif typefaces 40, 91, 237, 256
servers 265, 281, 285, 292, 304, 319, 320
service improvement 300
sexist language 222
sexual orientation discrimination 222, 223
sexual violence, images of 224
shapes (design) 42
shared material 264
short-answer questions 135
shortcut keys 93, 217
signal interference 268
signature recognition 282
simple sentences 226–7
single tables 60
site maps 47
skewed results 136, 160
slab serif typefaces 40
slack time 118, 160
sleep-on-it strategy 183
sliders 106
slideshow software applications 235, 236, 239
small-scale qualitative research 129
smartphones 232, 242, 263, 271
social constraints 37, 169, 200
social media 10, 12
software 37, 151–2, 220, 255, 266, 304
  copyright and 311–12
  data manipulation 234–5, 238–42
  hardware and 264
  piracy 310
  security 153, 275–81
  threats 272, 273, 291
  trial versions 273
solid state drives (SSD) 150, 218, 220, 256, 283
solutions 214
  constraints xiii, 168–70
  definition xiii
  design xiv, 37, 170
  development xiv–xv
  requirements xiii, 37, 52, 169–70, 176
  specifications 166–71
  testing 244–8, 253
sorting, testing 100, 108
sound data 18, 52, 136
sound muting 231, 232
source code corruption 273
source reputability 3
space design 175, 178
spam laws 170
spam servers 157
spatial design 38
spatial relationships 25
spellcheckers 88, 138, 240
spider diagrams 180
spoofing 139, 140
spreadsheets 19, 25, 26, 28, 30, 32–3, 52, 58, 60–1, 68, 70, 71,74, 82, 83, 84–8, 94–5, 137, 151, 169, 182, 191, 243
  checking 100
  design 84
  imported files and 80
  qualitative data 132–3
spyware 156, 273
SQL 19, 325
square brackets 15, 138
staffing concerns 134, 319, 322
standalone devices 263, 292
standards of service, workplaces 315
start date 123
static data visualisation 169
statistical analysis 3, 130
statistical infographics 195, 200
stereotype avoidance 222
still images 175
storage hardware 148, 149–50, 160, 283, 289
storyboards 47, 90
storytelling 220, 240
strategies, information management 298, 301–2
stream graphs 29, 52
streaming services 307
strings 17, 19, 74, 136
strong passwords 153–4, 302
Stuxnet worm 328
styles 137–8, 215
subfolders 216
sub-ideas 180
subjective data 125, 129, 130
subjective results 101, 168, 248, 249, 256
subscription download services 307
substitution 145, 182
subtitles 228
successor tasks 117, 160
sum field 85, 100, 102
summarising 137
summary statistics 100, 238
supporting data 125, 128, 168
surge protection 153
surveillance cameras 305
surveys 7, 52, 79, 130, 249, 315
  distortions 136, 160
  software 142
  tools 152
swipe cards 284
switches 151, 239, 266, 286
symmetric encryption 279
symmetrical balance 176
systems
  administrators 281
  analysts 300
  constraints 169
  GanttProject 121, 123
  goals 300, 334

9780170440875

layouts 192
protection software 280, 292
shutdown 327
software 151
tests 96

Tab key 77
Tableau 25, 94, 152, 209, 212, 220, 231, 234, 235, 238, 239
table of contents (TOC) entries 219
tables 31, 60, 68, 74, 108, 237
comma separation and 82
design 70–5
formatting 38
normalisation 60–6
validation 98, 99
tablet computers 232, 263, 265
tape backup 324
target audiences 10, 221–8, 234
tasks 169
changes to schedules 251
durations (GanttProject) 121–3
identification of 116, 120
management of 115
prioritisation 253
technical constraints 37, 169, 200
technical protocols 288–9, 292
technical terms 227, 229
technology mergers 287
telecommunications organisations 266
telephone line access 286–7
templates 152, 215, 236, 240, 252
test data 96, 97, 99, 108, 125, 248
testing xv, 52, 96–106, 178, 243, 244–8, 257
backups 325–6, 327
disaster plan 326–7
validation 25
testing tables 68, 90, 97–8, 101, 103, 108, 246–7, 248
text 91, 211, 237, 238
alignment 173–4
attributes 40
boxes 240
columns 175
data type (strings) 17, 18, 19, 52, 136, 137
editors 151, 242
files 79, 209
formatting 38
images on 174
location 189
replacement 229
size 228–30
strings 76
styles 39–40
testing 245
theoretical coding 131, 160
thermal printers 149
third normal form (3NF) 60, 65–6, 109
third person plural 223
third-party data storage 266, 290, 308, 312
threats, data and information 272–5, 292
time
allocation 118–19, 121–3
constraints 169
copyright (70 years) 306
data 27, 29, 30
timeframes 169, 195
timelines 118, 119, 193, 194, 242
constraints 233
data 23, 52, 93, 147
data recovery 332, 333
evaluation 250
project completion 254
solutions 249
testing 105
time saving backups 283
Times New Roman 40, 91
time stamping 18, 304
time visualisations 29, 30, 52, 193, 200
timing tests 248
tipping points 97, 245
titles (documents) 39–40
tool customisation 236
tooltip hover 212, 213
touchscreens 148, 173
trade secrets 304, 328
transaction processing systems 299
transborder data flows 312, 313
transfer errors 303
transgender audiences 222
transmission control protocol/internet protocol (TCP/IP) 288, 292
transmission media 263, 266
transparency, data reporting 145
transport layer security (TLS) 140. 281
tree charts 34
trojans 156, 273, 274
trust, qualitative survey 135
Twitter 10, 12, 13
type checks 24, 70, 109, 243, 257
typefaces 40, 91, 92, 102, 230, 233, 245

unambiguous data 145, 146
unauthorised users 314
uncompressed audio formats 18
unfair dismissal laws 315
unique identifiers 312
universal serial bus (USB) flash drives 150, 151, 283, 324
unnormalised tables 61
unpublished data 144
unrequested packets 280
unshielded twisted pair (UTP) cables 268–9, 292
unsolicited personal information 309
unsubstantiated suppositions 125
upselling 132
URLs 213, 240, 267
usability 38, 169, 170, 200
constraints 37
design 172–3
solutions 230, 234, 249, 250
tests 96, 106, 248
user acceptance tests (beta tests) 96, 101, 109
user error 93, 272, 292
user flow diagrams 193–4
user friendliness 172
user interface (UI) 46
usernames 273, 275–6, 292
user-productivity systems 299
utility software 151

validation 52, 70, 71, 72–3, 97, 108, 213, 243, 257
rules 58, 69–70, 79, 80, 88, 97, 101, 144, 147, 147
testing 80, 98, 99
vandalism 142
variables 127, 128, 209
Venn diagram 179–80
verification 243, 257
version control 108, 216, 217, 256, 264, 293
video conferencing 264
video files 233, 242
videos 89, 90, 92, 102, 175, 209, 238, 245
viral advertisements 140
virtual servers 267
viruses 156, 157, 273, 274, 280
vision statements 301, 300
Visual Basic 151, 242
visual information 211
visualisations 36, 95, 105, 183–4
visual media 91
visual weight 176
vocabulary 225–6, 245
voice recognition 282
voting trends (UK) 31

WAP/WEP encryption 270
warm backup sites 326
warning messages 101, 248
watermarks 140
weaknesses, primary and secondary data 5–6
weak passwords 154
weather data 9–10, 11–12, 22, 52
web-based software 152
web browsing, employees and 316, 317
web crawler 273
web links 189
weblogs (blogs) 251, 253, 257
Web Map Service (WMS) file 82–3, 109
webpages 237, 238, 239, 246, 267

9780170440875

web security protocol 281
web servers 267, 273, 277, 285, 289
websites 138, 155, 247, 273, 289, 300, 305, 308, 310, 311, 332
whiteboards 177, 179
white space 91, 92, 175, 176, 234
wide area networks (WANs) 266, 293
wi-fi 264, 265, 271
wind tunnels 193, 212
wiped files 219, 257
wired networks 267–9, 293
wire-framing 47, 48
wireless access point (WAP) 286, 293
wireless adapters 263, 285
wireless local area network (wi-fi) (WLANs) 151, 270
wireless networks 263, 264, 265, 270–1, 288, 293, 302
wireless security 156
word clouds 35, 52, 95
word frequency patterns 35
word processors 151
work breakdown structure (WBS) 116–17, 160, 254, 255
workplace email, privacy and 310
workplace responsibilities 315
worksheets 236, 242
workstations 265
World Wide Web Consortium (W3C) 231, 246
worms 156, 157, 274
Wozniak, Steve 185

XLM/XLMS file format 79, 82, 109

zoned security strategies 281, 293

9780170440875