Chapter 1
Data, information, knowledge and processing

Learning objectives

By the end of this chapter, you will be able to:

- define the differences between data, information and knowledge
- define static and dynamic data
- compare static information sources with dynamic information sources
- define direct and indirect data sources
- understand the advantages and disadvantages of gathering data from direct and indirect data sources
- understand how the accuracy, relevance, age, level of detail and completeness of information can affect its quality
- describe the coding of data and discuss its advantages and disadvantages
- evaluate the need to encode data and analyse different methods for encoding data
- define encryption and describe different methods of encryption
- evaluate the need for encryption and how it can be used to protect data
- discuss encryption protocols
- define validation and analyse a range of validation methods
- define verification and analyse verification methods
- explain the need for both validation and verification
- define proof reading
1.01 Data, information and knowledge

**KEY TERMS**

- **Data**: raw numbers, letters, symbols, sounds or images without meaning
- **Information**: data with context and meaning
- **Knowledge**: information to which human experience has been applied

**Data**

Data is raw numbers, letters, symbols, sounds or images with no meaning.

**EXAMPLE**

<table>
<thead>
<tr>
<th>Data</th>
<th>Context</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>P952BR</td>
<td>A product code</td>
<td>This is a product code but it is still not known what it is a product code for so it is still data.</td>
</tr>
<tr>
<td>@bbcclick</td>
<td>A Twitter handle</td>
<td>This is an address used for Twitter but it is not information unless it is known to be a Twitter handle or used within Twitter software. It’s also not known whose address it is.</td>
</tr>
<tr>
<td>359</td>
<td>Price in Pakistani Rupees</td>
<td>This is a currency value but it is not known what the price is for, so it is still data.</td>
</tr>
</tbody>
</table>

For the data to become information, it needs to be given meaning. Information is useful because it means something.

**EXAMPLE**

<table>
<thead>
<tr>
<th>Data</th>
<th>Context</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>P952BR</td>
<td>A product code</td>
<td>A product code for a can of noodles.</td>
</tr>
<tr>
<td>@bbcclick</td>
<td>A Twitter handle</td>
<td>The Twitter address for the BBC’s weekly technology show, <strong>Click</strong>, which is worth watching on BBC World News and BBC2 to keep up to date with technology.</td>
</tr>
<tr>
<td>359</td>
<td>Price in Pakistani Rupees</td>
<td>The price of a mobile phone cover.</td>
</tr>
</tbody>
</table>

Information

When data items are given context and meaning, they become information. A person reading the information will then know what it means.

Data is given context by identifying what sort of data it is. This still does not make it information but it is a step on the way to it becoming information as shown in the next example.

DISCUSSION POINT

When answering a question such as ‘Give one item of data’, do not try to explain what the data means because it then becomes information. Just give the raw numbers, letters, symbols or image.
Knowledge

Knowledge is basically what a person knows. This is known as their knowledge base. A knowledge base gets larger over time as a person gains experience or learning. Knowledge requires a person to understand what information is, based on their experience and knowledge base. Crown Prince Salman was appointed Crown Prince of Saudi Arabia on 18 June 2012. This is information. Knowing that he had been Crown Prince for 2 years on 1 August 2014 is knowledge. Knowledge allows data to be interpreted. In computing terms, knowledge is also what a machine knows through the use of a knowledge base consisting of rules and facts, often found in knowledge-based systems, modelling and simulation software.

### 1.02 Sources of data

#### KEY TERMS

- **Static data**: data that does not normally change
- **Dynamic data**: data that changes automatically without user intervention
- **Direct data source**: data that is collected for the purpose for which it will be used
- **Indirect data source**: data that was collected for a different purpose (secondary source)

#### Static data

Static means ‘still’. It is data that does not normally change. **Static data** is either fixed or has to be changed manually by editing a document.

#### Dynamic data

Dynamic means ‘moving’. It is data that updates as a result of the source data changing. **Dynamic data** is updated automatically without user intervention.

### Questions

A company creates websites using style sheets.

1. Identify one item of data that will be used by the company.
2. Describe how this item of data can become information.
3. Describe the term knowledge.

---

**Example**

<table>
<thead>
<tr>
<th>Information</th>
<th>Knowledge</th>
</tr>
</thead>
<tbody>
<tr>
<td>100km/h is the speed limit on expressways in Pakistan.</td>
<td>Travelling at 120km/h on expressways in Pakistan is illegal.</td>
</tr>
<tr>
<td>A red traffic light means a car should stop.</td>
<td>The price of a mobile phone cover is more expensive than a cup of coffee.</td>
</tr>
<tr>
<td>359 Pakistani rupees is the price of a mobile phone cover.</td>
<td>The price of a mobile phone cover is more expensive than a cup of coffee.</td>
</tr>
</tbody>
</table>

Figure 1.02 - A red traffic light.

Figure 1.03 - A good example of static data.

---

© in this web service Cambridge University Press

www.cambridge.org
TASK
Look at the Al Jazeera website www.aljazeera.com and identify which data is static and which data is dynamic (updates automatically).

### Static and dynamic information sources

<table>
<thead>
<tr>
<th>Static information source</th>
<th>Dynamic information source</th>
</tr>
</thead>
<tbody>
<tr>
<td>The information does not change on a regular basis.</td>
<td>Information is updated automatically when the original data changes.</td>
</tr>
<tr>
<td>The information can go out of date quickly because it is not designed to be changed on a regular basis.</td>
<td>It is most likely to be up to date as it changes automatically based on the source data.</td>
</tr>
<tr>
<td>The information can be viewed offline because live data is not required.</td>
<td>An internet or network connection to the source data is required, which can be costly and can also be slow in remote areas.</td>
</tr>
<tr>
<td>It is more likely to be accurate because time will have been taken to check the information being published, as it will be available for a long period of time.</td>
<td>The data may have been produced very quickly and so may contain errors.</td>
</tr>
</tbody>
</table>

**Table 1.01** - Static and dynamic information sources.

**DISCUSSION POINT**

Some people get a little confused by dynamic data because they think it can be any data that changes at any time. For example, some people think that any website includes dynamic data. However, if you look at www.aljazeera.com/contactus, the information on this page is mainly static, with links to other pages. Although it can be changed, it can only be done by changing the actual data on the page, whereas dynamic data changes because the original source has changed. If you look at www.aljazeera.com/programmes, you will see that most of the information on this page is sourced from programme schedules and programme information, which will be stored in another database. This makes it dynamic data because it will update when a new programme is scheduled.

**EXAMPLE**

- live sports results on a website (when a goal is scored, the scores will update on the website)
- news feeds on a mobile phone app (when the news is changed in the main database, the news feed will be updated on the phone)
- availability of tickets for a concert (when somebody books a ticket, the ticket is no longer available)
- product prices for a till/point of sale (if a price is reduced or increased in the database, this new price will be used the next time the barcode is scanned)
- train expected arrival times (these will update automatically based on the train’s position)
- profit for a product in a spreadsheet (profit = price – cost so when either the price or cost changes, then the profit changes too)

**TASK**

Look at the Al Jazeera website www.aljazeera.com and identify which data is static and which data is dynamic (updates automatically).
Direct data source

Data collected from a **direct data source** (primary source) must be used for the same purpose for which it was collected. It is often the case that the data will have been collected or requested by the person who intends to use the data. The data must not already exist for another purpose though. When collecting the data, the person collecting should know for what purpose they intend to use the data.

**EXAMPLE**

A sports shop wants to find out what other shops are charging for trainers. There are various direct sources that this data can be collected from. These could include:

- visiting the other shops and noting down the prices
- visiting the other shops’ websites and noting down the prices
- carrying out a survey of other shop owners to ask their prices (although they are unlikely to want to give these).

Indirect data source

Data collected from an **indirect data source** (secondary source) already existed for another purpose. Although it can still be collected by the person who intends to use it, it was often collected by a different person or organisation.

**EXAMPLE**

The sports shop could use various indirect sources to find out what other shops are charging for trainers including:

- carrying out a survey of customers who have purchased trainers from the other shops (the price was originally given to the customer for the purpose of selling the trainers which may have been given on a different date to when it is now being used or it may have been discounted at the time)
- looking at till receipts from the shop (the price is printed on the till receipt for the purpose of providing proof of purchase, not for identifying prices).

**TASK**

Which of the following are direct data sources and which are indirect data sources?

<table>
<thead>
<tr>
<th>Data</th>
<th>Reason collected</th>
<th>Reason used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Names and email addresses of members of a political party</td>
<td>To record their membership and to be able to contact them.</td>
<td>To contact members by email to see if they will donate some money.</td>
</tr>
<tr>
<td>Employee attendance dates and times</td>
<td>To identify when employees attended work and to calculate their wages.</td>
<td>To allow a police officer to check an employee’s alibi if a crime has been committed.</td>
</tr>
<tr>
<td>Flight times and prices from airline websites</td>
<td>To compare the prices and times for a trip to Florida.</td>
<td>To decide the best flight to use for a trip to Florida.</td>
</tr>
<tr>
<td>Names, ages and addresses of people</td>
<td>For a national census.</td>
<td>To allow a marketing company to find out which areas have the highest population of children.</td>
</tr>
<tr>
<td>Weather measurements from a weather station</td>
<td>To record the current weather.</td>
<td>To show the current temperature and rainfall on a website.</td>
</tr>
</tbody>
</table>
Advantages and disadvantages of gathering data from direct and indirect data sources

The general rule is that data collected directly for the purpose for which it is intended is more likely to be accurate and relevant than data that is obtained from existing data (indirect source).

<table>
<thead>
<tr>
<th>Direct data source</th>
<th>Indirect data source</th>
</tr>
</thead>
<tbody>
<tr>
<td>The data will be relevant because what is needed has been collected.</td>
<td>Additional data that is not required will exist that may take time to sort through and some data that is required may not exist.</td>
</tr>
<tr>
<td>The original source is known and so can be trusted.</td>
<td>The original source may not be known and so it can’t be assumed that it is reliable.</td>
</tr>
<tr>
<td>It can take a long time to gather original data rather than use data that already exists.</td>
<td>The data is immediately available.</td>
</tr>
<tr>
<td>A large sample of statistical data can be difficult to collect for one-off purposes.</td>
<td>If statistical analysis is required, then there are more likely to be large samples available.</td>
</tr>
<tr>
<td>The data is likely to be up to date because it has been collected recently.</td>
<td>Data may be out of date because it was collected at a different time.</td>
</tr>
<tr>
<td>Bias can be eliminated by asking specific questions.</td>
<td>Original data may be biased due to its source.</td>
</tr>
<tr>
<td>The data can be collected and presented in the format required.</td>
<td>The data is unlikely to be in the format required, which may make extracting the data difficult.</td>
</tr>
</tbody>
</table>

Table 1.02 - Direct and indirect data sources.

1.03 Quality of information

The quality of information is determined by a number of attributes.

Accuracy

Information that is inaccurate is clearly not good enough. Data must be accurate in order to be considered of good quality. Imagine being told that you need to check in at the airport 45 minutes before the flight leaves, so you turn up at 18:10 for a 19:05 flight only to find that you were actually supposed to check in one hour early.

Examples of inaccurate information include:

- decimal point in the wrong place, for example $90.30 instead of $903.00 could suggest a product is much cheaper than it really is
- misspelling such as ‘stair’ instead of ‘stare’, where words have completely different meanings
- misplaced characters, such as a licence plate of BW9EP3T instead of BW93PET.
Relevance

Information must be relevant to its purpose. Having additional information that is not required means that the user has to search through the data to find what is actually required.

**EXAMPLE**

Examples of irrelevant information include:

- being given a bus timetable when you want to catch a train
- being told the rental price of a car when you want to buy the car
- a user guide for a mobile phone that includes instructions on how to assemble a plug.

Age

Information must be up to date in order to be useful. Old information is likely to be out of date and therefore no longer useful. When using indirect data sources, always check when the information was produced.

**EXAMPLE**

Examples of out of date information include:

- the number of residents in a town based on a census from 2011, but 500 new homes have been built in the town since then
- a rugby score that has not been updated for 5 minutes during which time a player scored.

Level of detail

There needs to be the right amount of information for it to be good quality. It’s possible to have either too little or too much information provided. If there is too much information, then it can be difficult to find the exact information required. If there is not enough information, then it is not possible to use it correctly.

**EXAMPLE**

A person orders a pizza. They ask for a large pepperoni to be delivered. They forgot to say what type of base they wanted and where it should be delivered to. The pizza company does not have enough information to fulfil the order.

A traveller needs to catch a train from Bhopal to Kacheguda. They phone up to find out the time of departure and arrival, but they have to listen to all the times of the stations in between before they get the arrival time at Kacheguda.

Completeness

All information that is required must be provided in order for it to be of good quality. Not having all the information required means it cannot be used properly.

**EXAMPLE**

A person has booked their car in for a service over the phone. The mechanic tells them the name of the street but doesn’t give the building number.

**TASK**

Look at the invitation below.

Figure 1.08 - Birthday party invitation.

Come and Celebrate
Emmanuel’s
Birthday
11:00–1:30pm
18 Main Street
RSVP

There will be a magician. The magician was born on March 1st 1978 in Queen Elizabeth Hospital in Birmingham.

Describe how accuracy, relevance, level of detail and completeness affect the quality of information in the invitation.
1.04 Coding, encoding and encrypting data

KEY TERMS

Coding: representing data by assigning a code to it for classification or identification
Encoding: storing data in a specific format
Encryption: scrambling data so it cannot be understood without a decryption key to make it unreadable if intercepted
SSL: Secure Sockets Layer
TLS: Transport Layer Security
HTTPS: hypertext transfer protocol secure

Coding data

You are probably very familiar with coding data already. When you send a text message or instant message you probably code your data. You might use codes like:

• LOL = laugh out loud
• ROFL = roll on floor laughing
• FYI = for your information
• BTW = by the way
• 2 = to
• 4 = for
• BRB = be right back.

In a similar way, data stored on a computer can be coded. Coding is the process of representing data by assigning a code to it for classification or identification.

Often genders are coded as M for male and F for female. Clothing can be coded by the type, size and colour:

• DR = dress
• 2XL = extra extra large
• BL = blue
• DR2XLBL = a dress in size extra extra large and colour blue.

Using numbers for international dialling codes:

• 44 = Great Britain
• 33 = France
• 49 = Germany
• 34 = Spain
• 93 = Afghanistan
• 971 = United Arab Emirates
• 81 = Japan

Using abbreviations for international vehicle registration plates:

• GB = Great Britain
• F = France
• D = Germany (Deutschland)
• E = Spain (España)
• AFG = Afghanistan
• UAE = United Arab Emirates
• J = Japan
Advantages and disadvantages of coding data

There are a number of reasons for coding data. In the examples used above, it would take a long time to use all the letters of a country to dial a telephone number and there would not be enough space on the rear of a car to display the full country name.

The advantages of coding data can be summarised as shown in Table 1.03.

<table>
<thead>
<tr>
<th>Presentation</th>
<th>Data can be presented in small spaces, such as on labels or when listing large amounts of data.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Storage</td>
<td>Less storage space is required because less characters are used to store the data.</td>
</tr>
<tr>
<td>Speed of input</td>
<td>Data can be input more quickly because only a few numbers or letters need to be input instead of whole words or phrases.</td>
</tr>
<tr>
<td>Processing</td>
<td>Processors can process data more quickly because they only have to process the coded data instead of the whole word or phrase.</td>
</tr>
<tr>
<td>Validation</td>
<td>It's possible to validate data by checking it is a particular length or within a certain range or matching other rules. For example, an international vehicle registration code must be a maximum of three letters and not contain any numbers or symbols.</td>
</tr>
<tr>
<td>Confidentiality</td>
<td>Sometimes data can be coded so that it only makes sense to people who know what the codes mean.</td>
</tr>
<tr>
<td>Consistency</td>
<td>If the correct codes are used then data will be input consistently without spelling errors, making it easier to be searched or summarised.</td>
</tr>
</tbody>
</table>

Table 1.03 - Advantages of coding.

The disadvantages of coding can be summarised as shown in Table 1.04.

| Limited codes | There may not be enough codes available to use, meaning that more letters or numbers need to be added. For example, with international vehicle registration codes, E is already used for Spain and so Egypt has to use ET and Ecuador has to use EC. Now that Egypt has taken ET, Ethiopia can't use it and so has to use ETH. |
| Interpretation | Codes may be difficult to interpret. For example, with international vehicle registration codes, somebody might look at ET and assume it is Ethiopia or look at S and assume it is Spain. It's even more difficult when considering international dialling codes that are represented by numbers. |
| Similarity | Some letters and numbers can be difficult to distinguish such as O and 0 or Z and 2, especially if handwritten. With codes, it's difficult to guess what the value might be and so mistakes can occur. At least with words it's possible to guess what the badly written letter might be. |
| Efficiency | If a user inputting codes does not know what code to use, then they will not be able to enter the data efficiently. Instead they will have to look up the code. |
| Missing information | It's possible that some information gets lost during the process of coding. For example, if devices on a network are coded as L for laptop, P for printer, D for desktop computer and M for mobile phone, then information about whether or not the mobile phone or laptop is touch screen is lost. |

Table 1.04 - Disadvantages of coding.
Encoding data

When data is encoded, it is stored in a specific format. Computers do not recognise text, sound and images in the same way we do. Computers use binary digits which are 1s and 0s. One means on and zero means off. A typical binary number would look like this: 11011010. Therefore, data needs to be encoded into a format which the computer understands. Codecs are programs that are used to encode data for images, audio and video. The codecs are also needed to read the data.

Text

Text is encoded as a number that is then represented by a binary number. A common encoding method is ASCII (American Standard Code for Information Interchange). ASCII consists of 256 codes from 0 to 255. Here is a subset of the ASCII code table:

<table>
<thead>
<tr>
<th>Character</th>
<th>Decimal number</th>
<th>Binary number</th>
</tr>
</thead>
<tbody>
<tr>
<td>@</td>
<td>64</td>
<td>01000000</td>
</tr>
<tr>
<td>A</td>
<td>65</td>
<td>01000001</td>
</tr>
<tr>
<td>B</td>
<td>66</td>
<td>10000100</td>
</tr>
<tr>
<td>C</td>
<td>67</td>
<td>10000110</td>
</tr>
<tr>
<td>a</td>
<td>97</td>
<td>01100001</td>
</tr>
<tr>
<td>b</td>
<td>98</td>
<td>01100010</td>
</tr>
<tr>
<td>c</td>
<td>99</td>
<td>01100011</td>
</tr>
</tbody>
</table>

Table 1.05 - ASCII code table.

There are other encoding methods, too, such as extended ASCII which allows for more characters to be used, particularly for international languages. Other encoding methods include Unicode and EBCDIC but you don’t need to learn about these.

Images

Encoding is also used to store images. At the most basic level, images are encoded as bitmaps. A Microsoft Windows bitmap includes the following data when encoding an image (this is not the complete list):

- width of bitmap
- height of bitmap
- bit count which represents number of colours:
  - 1 = monochrome
  - 4 = 16 colours
  - 8 = 256 colours
  - 16 = 65536 colours
  - 24 = 16 million colours
- compression type (no compression, eight-bit run-length encoding or four-bit run-length encoding)
- horizontal resolution in pixels per metre
- vertical resolution in pixels per metre
- raster data (the actual shape of the image in pixels).

For the raster data, we will assume a monochrome image (black and white). Each bit will represent either a black or white pixel. A byte consists of eight bits and so will represent eight pixels. The encoding starts from the bottom left of the image, works to the right-hand side and then up to the next row and again from left to right.

TASK

Create a coding system for the names of people in your class. Show it to a friend. Can they interpret the code efficiently? Did you have any duplicate codes?

Visit the website www.ascii-code.com and look at the ASCII code table. What sentence does the following ASCII code represent?

084 104 101 032 098 105 103 032 098
097 100 032 119 111 108 102 033

Choose a language script and then choose a short sentence (less than ten words). Encode your sentence into UNICODE. Now show the Unicode to a friend and see if they can convert it back into the original sentence.