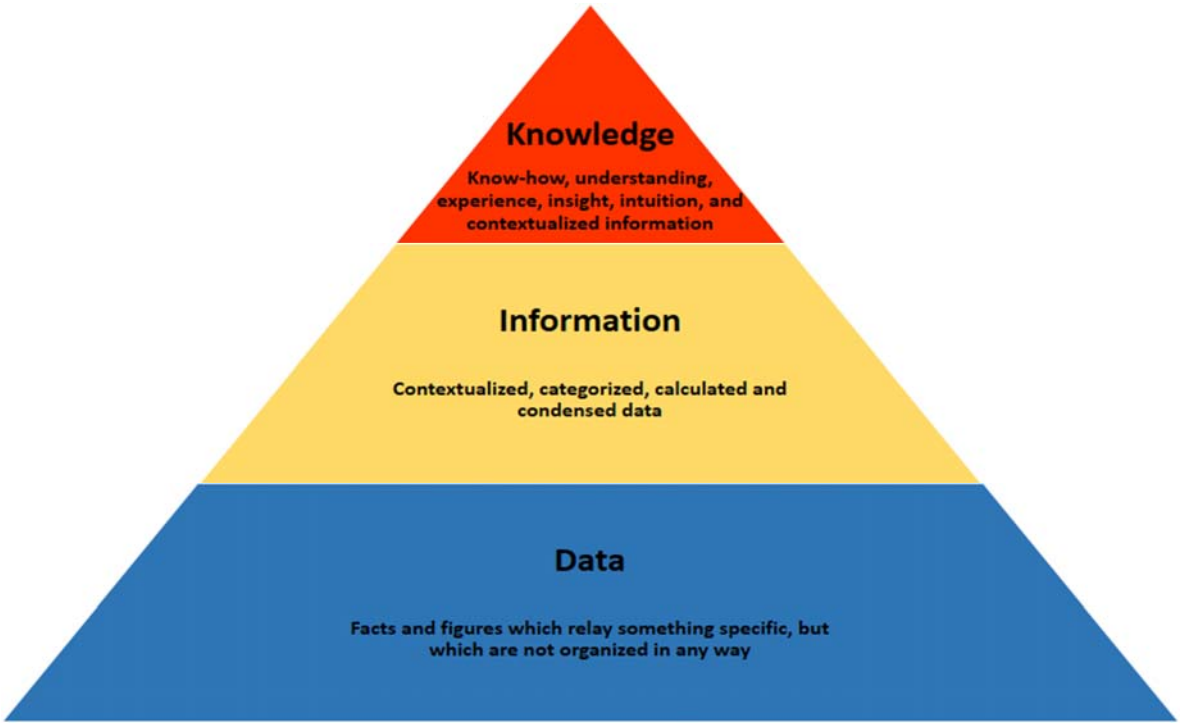
Data, Information and Knowledge

Data	Information	Knowledge
Data is comprised of the basic, unrefined, and generally unfiltered information	Information… is much more refined data… that has evolved to the point of being useful for some form of analysis	Knowledge resides in the user…happens only when human experience and insight is applied to data and information



- Data is unprocessed facts and figures without any added interpretation or analysis. "The price of crude oil is $80 per barrel."
- Information is data that has been interpreted so that it has meaning for the user. "The price of crude oil has risen from $70 to $80 per barrel" gives meaning to the data and so is said to be information to someone who tracks oil prices.
- Knowledge is a combination of information, experience and insight that may benefit the individual or the organization. "When crude oil prices go up by $10 per barrel, it's likely that petrol prices will rise by 2p per liter" is knowledge.

Sources of Data

Static and dynamic data

Difference between static and dynamic data is that once static data is created, the data that it contains doesn't change, whereas the data can change and update in dynamic data. An example of static data, is a newspaper, as once it has been printed, the information on it cannot be updated, whereas an example of dynamic data, would be a website, as that can be updated as and when needed.

Websites are classified as dynamic sources of data, because the information on the website can be constantly updated, for example if a sports website shows live football scores, it can be updated as and when goals have been scored therefore providing almost real-time information. Another reason why websites are classed as dynamic sources of data, is because users can interact with the site, e.g. If a website is that of football clubs, then a user can request new information about ticket availability for an upcoming game.

Situations where the use of static data would be preferable would be for a student trying to do homework, as the information needed for the work could be gained from a book, as that information would be reliable, and wouldn't change as it may have done if it were on a website. Another situation where static would be more preferable to dynamic data, would be if somebody wanted to gain information from a CD, or book, as it would be accurate, as that would've gone through a lot of editing before being released.

Problems with using dynamic data are that sometimes it may not be as accurate as static data. For example, information on a website isn't necessarily accurate as anybody can create a website. Another problem with dynamic data is that if somebody needed to reference something for example on a website, the information on that website could've changed, or been deleted since they last checked.

Static Information Sources

Advantages	Disadvantages
Usually checked for accuracy so often more reliable than dynamic sources	If held on a CD, the disk could be lost or damaged
Can be used without access to the internet.	Can take a long time to produce i.e. in terms of checking, publishing and distribution
If the information is paper-based then it can be used without access to a computer.	Specialised information can be expensive to purchase
Have a historical copy to refer back to even if a later version is produced. Whereas with a dynamic source, the data can be overwritten or vanish	May have a licence agreement which means that only one person can use it at a time and that copies cannot be made
Can be distributed to the people who have a specific interest in the information e.g. customers, teachers	If a mistake is found in the information then new copies need to be printed and distributed
	Limited to the information available in the CD/book - not so easy to cross-reference as it would be with a web site.
	If in a CD/DVD format then will need access to a computer.

Dynamic Information Sources

Advantages	Disadvantages
Can be updated very quickly as changes happen	Changes can be made quickly and mistakes may go unnoticed
Often fairly inexpensive to maintain when compared with producing static information sources	You may need certain hardware such as a computer to access the internet or a mobile phone to access information services.
A large number of people can access the information at the same time	There may be conflicting information on different web sites
It is easy to find related information or to cross-reference information	The information may be biased or completely wrong
You don't need to know where to look for the information because you can use a search engine to locate relevant web pages	Information overload - it can feel overwhelming if you don't know what you are looking for.
Can access a lot of free content	

Direct and indirect data sources

Direct data:
Data that has been collected from an original source.

Indirect data:
Data that has been used for a purpose different to the purpose as to why it was collected in the first place. People/companies involved in collecting the data are different to those using the data, e.g, market surveys that sell the results to other companies.

	Advantages	Disadvantages
Direct	• The source and collection method is known and verified. • The exact data required can be collected. • The data being collected can change in response to answers.	• May not get a large range of data. • Data may not be available – location/time.
Indirect	• Large range of data is available that could not have been collected directly. • Data can be available from different locations and time periods. • Analysis might already have been completed on some of the data.	• Do not know if any bias was placed on the collection. • Cannot be certain of accuracy of the recording of data. • May not have all the information about how, when and where it was collected to make a valued opinion on its usefulness. • If the information was not originally collected, may not be able to get hold of it.

Factors that effect the quality of information

ACCURACY:
The data that has been collected must be accurate, otherwise the information it will produce will be inaccurate.

RELEVANCE:
In order for information to be useful, data must be relevant.

AGE:
In order for the information to be useful, the data needs to be up to date. Information changes over time, so old, out-of-date information can be misleading.

COMPLETENESS:
In order for information to be useful it needs to be complete. If parts of information are missing then you will not be able to make use of it or make accurate decisions.

PRESENTATION:
Information that is presented in a disorganized way or manner that is hard to understand will be less useful to you and of little value. Sorting or organizing data before you present it can make it easier to understand and be more useful.

LEVEL OF DETAIL:
Giving too much information will make it difficult to find what you require. Whereas, too little information will make it hard for you to understand or make use of the information provided.

Encoding data

Encoding is the process of converting data from one form to another. While "encoding" can be used as a verb, it is often used as a noun, and refers to a specific type of encoded data. There are several types of encoding, including image encoding, audio and video encoding, and character encoding.

Media files are often encoded to save disk space. By encoding digital audio, video, and image files, they can be saved in a more efficient, compressed format. Encoded media files are typically similar in quality to their original uncompressed counterparts, but have much smaller file sizes. For example, a WAVE (.WAV) audio file that is converted to an MP3 (.MP3) file may be 1/10 the size of the original WAVE file. Similarly, an MPEG (.MPG) compressed video file may only require a fraction of the disk space as the original digital video (.DV) file.

Character encoding is another type of encoding that encodes characters as bytes. Since computers only recognize binary data, text must be represented in a binary form. This is accomplished by converting each character (which includes letters, numbers, symbols, and spaces) into a binary code. Common types of text encoding include ASCII and Unicode.

Whenever data is encoded, it can only be read by a program that supports the correct type of encoding. For audio and video files, this is often accomplished by a codec, which decodes the data in real-time.

Codec

The name "codec" is short for "coder-decoder," which is pretty much what a codec does. Most audio and video formats use some sort of compression so that they don't take up a ridiculous amount of disk space. Audio and video files are compressed with a certain codec when they are saved and then decompressed by the codec when they are played back. Common codecs include MPEG and AVI for video files and WAV and AIFF for audio files. Codecs can also be used to compress streaming media (live audio and video) which makes it possible to broadcast a live audio or video clip over a broadband Internet connection.

.WAV files

Standard digital audio file format used for storing waveform data; allows audio recordings to be saved with different sampling rates and bitrates; often saved in a 44.1 KHz, 16-bit, stereo format, which is the standard format used for CD audio.

.MP3 files

Compressed audio format developed by the Moving Picture Experts Group; uses "Layer 3" audio compression; commonly used to store music files and audiobooks on a hard drive; may provide near-CD quality sound (stereo, 16-bit) in a file roughly 1/10 the size of a .WAV or .AIF file.

The quality of an MP3 file depends largely on the bit rate used for compression. Common bit rates are 128, 160, 192, and 256 kbps. Higher bit rates result in higher quality files that also require more disk space.

.MPEG

Common digital video format standardized by the Moving Picture Experts Group (MPEG); typically incorporates MPEG-1 or MPEG-2 audio and video compression; often used for creating movies that are distributed on the Internet.

Coding Data

What is coding of data?

Any system will need to have data collected, entered and stored. One method of storing data is to **assign codes** to it. This usually means shortening the original data in an agreed manner.

Example 1

Original data: Monday; Tuesday; Wednesday; Thursday; Friday

Coded data: Mon; Tues; Wed; Thurs; Fri

Example 2

Original data: Xtra Large; Large; Medium; Small

Coded data: XL; L; M; S

Reasons to code data:

It is common for much of the data collected and entered into a system to have some degree of repetition and redundancy i.e. extra information that does not add anything. And this pattern or repetition is why it is efficient to code the data in some way.

- **Speeding up data entry**

  Let's take the example of collecting data about a person's gender. People can be either 'Male' or 'Female'.

  Whilst these two options are easily understood by all, imagine having to enter the word 'Male' and 'Female' into a system many hundreds of times. It is a waste of time and effort because no extra information is contained in the full words compared to a single letter.

- **Increase accuracy of data entry**

  The other issue is that no matter how accurate a person is at data entry, at some stage they are likely to make a mistake and might spell 'Male' as 'Mail' or 'Female' as 'Femal'. This type of mistake will make any results from your database queries unreliable.

  Instead of entering 'Male' or 'Female' you could code the data and instead enter it as 'M' or 'F'.

  Simply having to enter one letter instead of a possible six will speed up data entry. It will also cut down on the risk of mistakes being made with spelling.

- **Use of validation**

  When data has been coded it makes it easier to use validation to check if the data entered is sensible. With the example above, the person entering the data could still make a mistake and enter 'S' instead of 'M' or 'F'.

  But if you set up validation so that the field will only accept the letters 'M' or 'F' and absolutely nothing else then that should further cut down on possible mistakes.

- **Less storage space required**

  Every letter that you store in your database system will take at least one byte of storage. If you store 'Female' as 'F' then you will save five bytes of storage space. If the system belongs to a large organization, there might be many thousands or millions of records stored - simply by coding one field, a huge amount of hard disk storage can be saved.

- **Faster searching for data**

  The smaller the size of your database, the faster it will be to search and produce results. Thus by coding data and keeping the size of the system to a minimum the more time you can save in the long run when running queries.

Problems caused by coding data:

Whilst coding data can bring many benefits it can also lead to some problems.

- **Coarsening of data**

  This means that during the coding process some of the subtle details in the data are lost.

  The colours could be classed as:

  Light pink, pale blue, black and mid blue

  However, when these colours are coded they may become:

  PK (pink), B (blue), BK (black), BE (blue)

  In this case, no allowance has been made for shades of colours. The fine detail has been lost. This is what is meant by 'coarsening of data'.

- **Coding can obscure the meaning of the data**

  A reader seeing the 'gender' data as M/ F is pretty likely to know that it means Male/ Female.

But some codes are more obscure, for example the country code for Switzerland is CHE. Many people might not recognize what this code represents.

- **Coding of Value Judgments**

  When you are collecting data about people's opinions it might be difficult to code their answers with accuracy. The code they give will depend on their individual opinion. Coding of value judgments will inevitably lead to coarsening of the data since there will be a wide range of opinions that could be held and only a limited number of codes available.

Further examples of data coding:

In our everyday lives we come across many examples of how coding is used to represent data. Here are just a few more ideas:

- **Country names**

  The name of a country can be represented by two letters. For example:

  Great Britain - GB

  France - FR

  Canada – CA

- **Airline flight codes**

  When you fly you may have noticed that your flight is given a code.

  This code consists of two letters to identify the airline that you are flying with. The letters are usually followed by numbers to represent a particular route.

  Examples:

  So for example, a British Airways flight from Heathrow to Oslo might be coded as BA766.

  A flight operated by the airline company Emirates which depart from Dubai and arrives at Heathrow might be coded as EK029.

Advantages of coding:

- Data entry can be faster

- Data entry can be more accurate

- Validation can further improve accuracy

- Less storage space required

- Faster searching for data

- Coded data can be more secure if people don't know what it means

Disadvantages of coding:

- Coarsening of data

- Meaning of data can be obscured

- Value judgments are difficult to code

- If people don't know the code it can slow down data entry

- If codes are complicated they might be entered incorrectly

- Might run out of code combinations

Data Validation and Verification

Validation:

Validation is one way of trying to reduce the number of errors in the data being entered into your system. Validation is performed by the computer at the point when you enter data. It is the process of checking the data against the set of validation rules.

Validation aims to make sure that data is sensible, reasonable, complete and within acceptable boundaries.

Data validation can be performed by using a number of validation checks.

Range Check

A range check is commonly used when you are working with data which consists of numbers, currency or dates/times.

A range check allows you to set suitable boundaries:

Boundary	Description	Validation
Upper limit	The maximum price of any item in a shop is £100	<=100
Lower limit	In a shop, you cannot sell a negative number of items, however you can sell no items	>=0
A range	to achieve a B grade you must score between 75% - 84%	>=75 AND <=84

Type Check

When you begin to set up your new system you will choose the most appropriate data type for each field.

A type check will ensure that the correct *type* of data is entered into that field. For example, in a clothes shop, dress sizes may range from 8 to 18. A number data type would be a suitable choice for this data. By setting the data type as number, only numbers could be entered e.g. 10, 12, 14 and you would prevent anyone trying to enter text such as 'ten' or 'ten and a half'.

Check Digit

This is used when you want to be sure that a range of numbers has been entered correctly. There are many different schemes (algorithms) for creating check digits.

For example, the ISBN-10 numbering system for books makes use of 'Modulo-11' division. In modulo division, the answer is the remainder of the division. For example

8 Mod 3 = 2 i.e. the remainder of dividing 8 by 3 is 2.

Consider the ISBN number:

ISBN 1 84146 201 2

The check digit is the final number in the sequence, so in this example it is the final '2'.

The computer will perform a complex calculation on all of the numbers and then compare the answer to the check digit. If both match, it means the data was entered correctly.

Length Check

Sometimes you may have a set of data which always has the same number of characters.

For example a UK landline telephone number has 11 characters.

A length check could be set up to ensure that exactly 11 numbers are entered into the field. This type of validation cannot check that the 11 numbers are correct but it can ensure that 10 or 12 numbers aren't entered.

A length check can also be set up to allow characters to be entered within a certain range.

For example, postcodes can be in the form of:

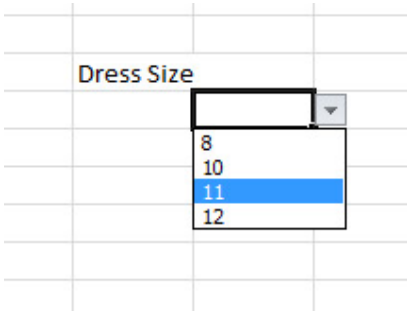
CV45 2RE (7 without a space or 8 with a space) or

B9 3TF (5 without a space or 6 with a space).

So you could set a length check for postcode to accept data which has a minimum number of 5 characters and a maximum number of 8.

Lookup Check

Where you have a field which only allows a limited list of items to be entered then a lookup list can help to reduce errors.



For instance, the image opposite shows a 'look-up' list being used in a spreadsheet that only allows 8, 10, 11 or 12 to be entered.

For example:

- a shop might put the dress sizes into a lookup list

- a car showroom might put the car models into a lookup list

- a vet might list the most popular types of animals that they deal with

Picture/Format Check

You may see this validation technique referred to as either a picture or a format check, they are the same thing.

Some types of data will always consist of the same pattern.

Example 1

Think about a postcode. The majority of postcodes look something like this:

CV36 7TP

WR14 5WB

Replace either of those examples with L for any letter which appears and N for any number that appears and you will end up with:

LLNN NLL

This means that you can set up a picture/format check for something like a postcode field to ensure that a letter isn't entered where a number should be or a number in place of a letter.

Example 2

A National Insurance number must be in the form of XX 99 99 99 X. The first two and the last characters must be letters. The other six characters are numbers. Any format entered differently to this will be rejected.

Presence Check

There might be an important piece of data that you want to make sure is always stored.

For example, a school will always want to know an emergency contact number, a video rental store might always want to know a customer's address.

A presence check makes sure that a critical field cannot be left blank, it must be filled in. If someone tries to leave the field blank then an error message will appear and you won't be able to progress to another record or save any other data which you have entered.

Verification:

Verification means to check that the data on the original source document is identical to the data that you have entered into the system. Verification can be performed in two ways; double entry method, visual check.

Double entry

Think about when you choose a new password, you often have to type it in twice. This lets the computer check if you have typed it exactly the same both times and not made a mistake. It verifies that the first version is correct by matching it against the second version.
Whilst this can help to identify many mistakes, it is not ideal for large amounts of data.

- It could take a person a lot of time to enter the data twice.
- They could enter the same mistake twice and so it wouldn't get picked up.
- You would end up with two copies of the data.

Visual check

This saves having to enter the data twice. It can help pick up errors where data has been entered incorrectly or transposed.

However, it isn’t always that easy to keep moving your eyes back and forth between a monitor and a paper copy. Also, if you are tired or your eyes feel 'blurry' then you might miss errors.

Data Encryption

What is Encryption?

Encryption means to scramble data in such a way that only someone with the secret code or key can read it.

Why is it important?

Today, encryption is far more sophisticated, but it serves the same purpose - to pass a secret message from one place to another without anyone else being able to read it.

Encryption is extremely important for e-commerce as it allows confidential information such as your credit card details to be sent safely to the online shop you are visiting.

Web browsers are able to encrypt your purchase details using an encryption method called 'SSL' (Secure Socket Layer). You know this is switched on when a small padlock appears in the bottom right of the browser. SSL gets switched on when you visit a 'secure server' that has an address that starts with HTTPS:// (note the 'S').

How does it work?

Encryption works by scrambling the original message with a very large digital number (key). This is done using advanced mathematics. Commercial-level encryption uses 128 bit key that is very, very hard to crack. The computer receiving the message knows the digital key and so is able to work out the original message.

Problems with encryption

There are three problems;

a) It is slower than normal browsing. It takes a while for the browser to do the maths required to scramble the message and another delay on the server that has to unscramble the data.

b) Online shops have to have a digital certificate that contains part of the key. This is not free and has to be supplied by a 'certificate authority'.

c) It can be a complicated business running a secure server, so very often, ordinary online shops will hire a specialist 'Payment Gateway' such as WorldPay or Paypal to handle payments for them.

Symmetric vs Asymmetric encryption

Symmetric Encryption

Symmetric encryption's job is to take readable data, scramble it to make it unreadable (protecting it from prying eyes while it's being stored on a disk or transmitted over a network), then unscramble it again when it's needed. It's generally fast, and there are lots of good encryption methods to choose from. The most important thing to remember about symmetric encryption is that both sides—the encrypter, and the decrypter—need access to the same key.

Asymmetric Encryption

Asymmetric encryption also takes readable data, scrambles it, and unscrambles it again at the other end, but a different key is used for each end. Encrypters use a public key to scramble the data, and decrypters use the matching private (secret) key on the other end to unscramble it again.

The public key means that it can and should be published. (This is why asymmetric encryption is also often referred to as public-key encryption), but the private key must be kept private, protected much like the key for symmetric encryption.