Component 1: Data representation and storage

Key terms

Term	Definition
Bit	A binary digit 1 or 0.
Metadata	A set of data that gives information about other data.
Character	A letter, digit, space, punctuation mark or various other keyboard symbols.
Character set	A table that maps a character with a unique binary number.
Compression	The process of making a file size smaller.
Array	A data structure that can hold a fixed number of data items, which must be of the same data type.

Storage requirements

Unit	Symbol	Value
Bit	b	1 bit
Nybble	-	4 bits
Byte	B	8 bits
Kilobyte	kB	1024 bytes
Megabyte	MB	1024 kB
Gigabyte	GB	1024 MB
Terabyte	TB	1024 GB
Petabyte	PB	1024 TB

Calculating storage requirements

File sizes are calculated as follows:

Graphics: *width* × *height* × *colour depth*

Sound: *channel* × *sample rate* × *bit depth*

Metadata

Examples of metadata used in graphics and sound files:

- Genre: the genre that the sound file belongs to
- Artist: the name of the artist who sang the songs
- Date Created / Year: the date the graphic was taken
- Location: the location where the graphic was taken
- Colour depth
- Dimensions – the width and height of a graphic.

Representation of characters

Characters are stored on a computer system as a binary number using a character set. Examples of character sets include ASCII and Unicode.

A small part of the ASCII character set:

Denary	Binary	Hex	Character
64	1000000	40	@
65	1000001	41	A
66	1000010	42	B
67	1000011	43	C

Character sets allow for meaningful data to be exchanged between different computer systems.

Unicode has combined and replaced many others character sets. It was originally an extension to the ASCII character set, but now contains many of the characters used around the world. Each character requires 2 bytes of storage.

Compression

Lossless compression

Uses an algorithm that compresses data into a form that may be decompressed without any loss of data (recovers all original data).

Lossless data compression works by finding repeated patterns in data and compressing those patterns in an efficient manner.

For this reason, lossless data compression is also referred to as redundancy reduction.

Lossless data compression is ideal for text.

Lossy compression

Compressed files can never be recovered exactly as they were before they were compressed. When compressed files are decompressed they do not give back the original data, i.e. data is lost.

Because lossy compression cannot be decompressed to yield the exact original data, it is not a good method of compression for critical data, such as textual data.

Lossy compression is used to compress multimedia data, such as images, sound and video, for internet streaming.

Calculating compression ratios

$$\textit{Compression ratio} = \frac{\textit{Original file size}}{\textit{Compressed file size}}$$