Data Representation

Computers and data

Computers store and process data. Every item of data is stored as a series of numbers.

Every instruction in a program is also stored as numbers called **machine code**. A program is converted (or **compiled**) into machine code before being run.

Encoding data

As a result, any piece of data needs to be **encoded** as a set of numbers. We can write these numbers down as decimal (or denary) numbers), but in the end the computer will turn them into binary.

Encoding text characters

Words are made up of **characters**. These can be letters but can include other characters - such as punctuation marks, numbers or even spaces.

To store words in a computer we need a way of encoding each character as a number. This is called a **character code**. A complete collection of characters is a **character set**.

Every word is made up of symbols or characters. When you press a key on a keyboard, a number is generated that represents the symbol for that key.

ASCII Character Set

The **ASCII character set** was first used in 1963. It was developed as a way of encoding characters so they could be sent between teletype machines. These allowed text to be sent quickly along telephone wires and printed at the other end.

ASCII stands for American Standard Code for Information Interchange

ASCII code allows the characters which appear on a standard keyboard to be encoded. Code 065 represents 'A', 094 represents '^' and 032 represents a space. This means that standard English language messages to be sent.

A teletype machine

ASCII code has 128 different characters. This allows it to be represented in binary code by **7 bits**. This allowed messages to be sent quickly enough to be useful and not take up too much memory. A 7 bit binary number allows numbers up to 127 to be stored - so, including 0, there are 128 different codes possible.

A binary **bit** is one binary number - either a 1 or a 0. 7 bits means that a sequence of seven binary digits are used in a row.

ASCII characters are grouped so that a human user can use logic to work through them more effectively.

The limitations of ASCII

ASCII code is limited to 128 characters. It's not difficult to think of cases where you need access to more character than that - adding an accent to an e in French, for example, or representing the Danish character Ø. And that's without adding Arabic or Chinese characters.

As computers became more powerful and communication spread across the world, ASCII's limit on characters soon became a problem. At the same time, communication systems became quicker so there was no longer a need for each character to be limited to seven bits.

The solution – Unicode

This led to the development of **Unicode** as an expanded character encoding system.

Unicode retains the exact same sequence as ASCII for the first 128 characters. It then adds many more codes to represent other characters.

Every major language used today is included in Unicode, as well as symbols used in areas such as Mathematics (for example, Pi and ∑). It also includes codes which developers can use to define their own symbols.

Unicode is meant to be a UNIversal character set. Work on it began in 1987. It is still growing - over 7,000 characters were added in June 2015.

> Unicode uses between 8 and 32 bits per character, so it can represent characters from languages from all around the world.

> Unicode includes symbols such as playing card symbols, musical notes and Egyptian hieroglyphs. In 2010 emoticons and emojis were added.

Activities:

1. Look at the ASCII character table.
    a) What range of numbers represent the lower-case letters?
    b) What rage of numbers represent the upper-case letters?
    c) What range of numbers represent digits?
2. a) Explain how far you agree with this statement: "ASCII characters are grouped together in logical groups"
    b) Why do you think the pattern of ASCII characters is used?
3. Explain why a set of agreed codes for standard characters were needed.
4. Suggest two disadvantages to using the ASCII character set.
5. Why do you think the first 128 characters ASCII characters were retained in the same order in Unicode?
6. Explain why it was necessary to develop Unicode.

Key Vocab:

Character set   **Character code**   **Bit**   **ASCII**   **Unicode**   **Encoding**