Text Representation – ASCII and Unicode

Computers and data

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

Any piece of data stored in a computer has to be turned into numbers – and eventually end up as binary.

This means that text like this needs to be stored as numbers.

Anything you type on a keyboard needs to be **encoded** as a set of numbers. We can write these numbers down as **decimal** numbers, but in the end the computer will turn them into **binary**.

Encoding text characters

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.

Words are made up of individual **characters**. These can be letters but include other characters –punctuation marks, numbers and 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 character codes is called a **character set**.

Once you have a character set that everyone agrees on, computers can communicate with each other. The first major character set developed for use in computers was called **ASCII**.

ASCII Character Set

ASCII was developed in America. It stands for American Standard Code for Information Interchange

The **ASCII character set** was first used in 1963. It was developed in America 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.

Computers had very limited memory and communication links were much slower then they are today. This meant that keeping the number of bits to store each character low was a real advantage.

ASCII code uses only the characters which appear on a standard keyboard. These are the only characters which are included. Each character has a number to represent it: code 065 represents 'A', 094 represents ^ and 032 represents a space.
This means that standard English language messages can be sent.

A teletype machine

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 code only has **128 different characters**. This allows it to be represented in binary code by **7 bits**.

Using 7 bits allowed messages to be sent quickly enough to be useful and not take up too much memory. A 7 bit number allows numbers up to 127 to be stored – so, including 0, there are 128 different codes.

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.

Because ASCII was developed in America it uses the standard American-English alphabet and symbols – which means that there's no pound sign…

The solution – Unicode

The need to add more symbols led to the development of **Unicode** as an expanded character set.

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, π and ∑). It includes codes which developers can use to define their own symbols.

Unicode is meant to be a **Universal** character set that would work for any language. Work on it began in 1987. It is still growing – over 7,000 characters were added in June 2015.

Unicode uses up to 32 bits per character, so it can represent characters from many languages. As of September 2022 Unicode included codes for 149,186 characters.

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

Activities:

1. What does the term **character** mean?
2. How is each character stored in a computer?
3. What is a **character set**?
4. **ASCII** is an example of a character set:
   a) When was ASCII first used?
   b) Where was ASCII developed?
   c) How many character codes are included in the ASCII character set?
   d) How many binary bits are used to store each character code?
   e) Give two reasons why using this many bits was an advantage originally?
5. Look at the ASCII character set:
   a) What range of numbers represent the lower-case letters?
   b) What range of numbers represent the upper-case letters?
   c) What range of numbers represent digits? (0, 1, 2 etc…)
   d) Why do you think the pattern of ASCII characters is used?
6. ASCII has now been replaced by **Unicode**:
   a) What are the advantages of Unicode over ASCII?
   b) How was it possible to increase the number of bits used to store each character?
   c) Why do you think the first 128 ASCII characters were retained in the same order in Unicode? What technical advantages would this have?