Data Representation

Storing Text

Learning Intentions

By the end of this lesson you will be able to:

- Explain what **Unicode** is
- Describe the advantages of Unicode over **ASCII**
- Describe the storage requirements for Unicode
- Describe how many characters can be represented by ASCII

A question…

Computers can only use and understand binary digits which are 0's and 1's.

How can we store text?
By storing them as binary numbers….

ASCII Codes

Each character on a keyboard has its own **ASCII** code. This is a binary value that represents each character that can be seen on the screen.

American **S**tandard **C**ode for **I**nformation Interchange

For example A = 0100 0001 (65)

Although ASCII originally was a 7 bit code which could represent 128 characters

IBM used 8 bit units.

This became known as **extended ASCII** and formed the basis of the ISO 8859 standard.

This allowed additional characters to be stored such as © and ™

ASCII Table

An extract of the ASCII Table is shown below:
The ASCII table has 128 values

52 just for text. 10 for numbers

Spacebar and tab key have codes too

What about the rest?

Code	Symbol	Code	Symbol	Code	Symbol	Code	Symbol
48	**0**	**78**	**N**	**64**	**@**	**97**	**a**
49	**1**	**79**	**O**	**65**	**A**	**98**	**b**
50	**2**	**80**	**P**	**66**	**B**	**99**	**c**

Control Characters

The rest of the ASCII code are reserved for **control characters**.

These are non-printable characters that have an effect such as

The trusted delete/backspace keys?

Character Set

The **character set** is the name given to the complete set of characters that the computer can represent.

Different character sets are used to represent different languages

The character set can alter the layout of the keyboard
For example on American layout keyboards the @ sign is above the number 2 not the “ mark.

Summary

Text is stored in the computer using ASCII values
1 ASCII Value = **7 bits** of memory
Extended ASCII = **8 bits**

Control characters are the **non-printable characters** which have an effect on the screen such as the Enter Key

The **character set** is the name given to the entire set of characters that the keyboard can produce

Unicode

A little more about ASCII

Although **ASCII** originally was a **7 bit code** which could represent 128 characters

IBM used 8 bit units.

This became known as **extended ASCII** and formed the basis of the ISO 8859 standard.

This allowed additional characters to be stored such as **© and ™**

We have a problem with ASCII

You are planning a trip to the far east and want to learn some of the language
Think of a simple phrase such as “hello”
Although in English this would use 5 bytes using ASCII

In Chinese it is: **Nǐ hǎo**

The symbol for this is 你好

This is **impossible** to represent using the ASCII code.

Unicode is an alternative method used to represent text.

Unicode

The Unicode® or Universal Character Set(UCS) standard was originally a 16 bit code

Version 1.0.0 was released in 1991 containing approx. 7000 c

The most common encodings are Unicode Transmission Format (UTF)

- UTF16 and UTF 8

At present it can require 8,16 or 32 bits

This allows it to **represent** far more characters and languages.

It does however mean that it takes up more memory space to store a single character.

Summary

Unicode® will usually use 8/16 bits per character

Means it can represent more characters/languages than ASCII

Current standard has over 120,000 characters

But this does also means it takes up more memory storage space per character