Information Management Resource Kit

Module on Management of Electronic Documents

UNIT 2. FORMATS FOR ELECTRONIC DOCUMENTS AND IMAGES

LESSON 7. CHARACTER ENCODING: LATIN AND NON-LATIN SCRIPTS

NOTE

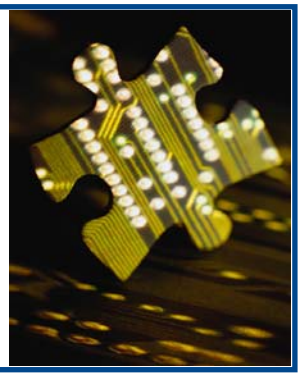
Please note that this PDF version does not have the interactive features offered through the IMARK courseware such as exercises with feedback, pop-ups, animations etc.

We recommend that you take the lesson using the interactive courseware environment, and use the PDF version for printing the lesson and to use as a reference after you have completed the course.

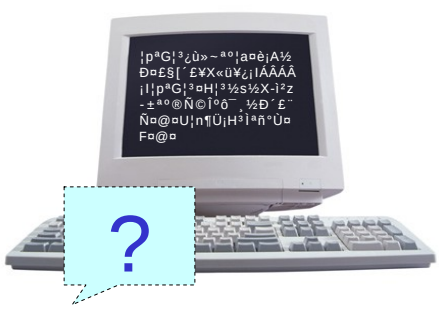
Objectives

At the end of this lesson, you will be able to:

• understand how to solve the main problems with character **encoding**.



Character encoding

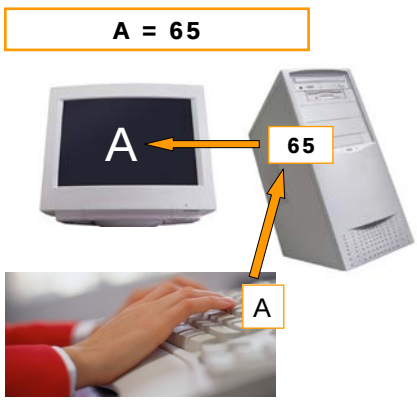


Probably you have at some time or another opened a web page only to find illegible text and meaningless characters.

This is a problem connected to **character encoding**.

Let's try to understand the causes of this problem…

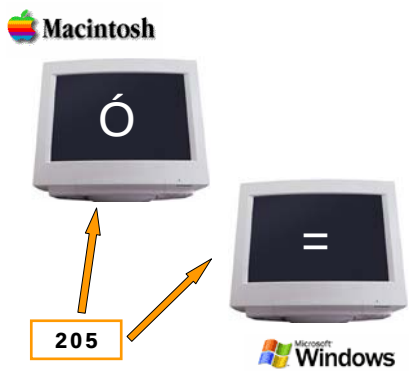
Character encoding



Character encoding is the organization of a set of **numeric codes** that represent all meaningful characters (single letter, digit, space, punctuation, etc.) of a script system in memory.

Each **character** is **stored** in memory **as a number**.
When a user enters characters, the user's keypresses are converted to character codes; when the characters are displayed on screen, the character codes are converted to the glyphs of a font.

Character encoding



In most character encoding standards, the character set changes to represent the language being used, so the upper-level characters may include symbols, accented Roman letters, Cyrillic, or other characters, **depending on the character encoding chosen**.

For example, the character "Ó" in the Macintosh Standard Roman Character Set is in the same code point 205 as the equal sign "=" in Windows extended ASCII encoding.

Character encoding

Several encoding systems are available for which encoding schemes have been developed:

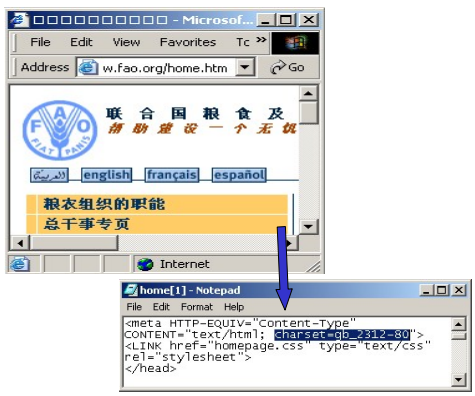
	7-BIT ENCODING SYSTEM	8-BIT ENCODING SYSTEM	16-BIT ENCODING SYSTEM	32-BIT ENCODING SYSTEM
What is	An encoding system that uses a fixed width of 7-bit encoding that allows for a character set of **128** values (2^7).	An encoding system that uses the eighth bit (parity bit) of the 7-bit encoding system to cover a larger number of characters. It allows for the use of **256** values (2^8).	An encoding system that uses a fixed-width of 16 bits per character, which allows the accommodation of a total of **65536** (2^{16}) values.	Standard named ISO/IEC 10646-1. It is essentially a 31-bit encoding, i.e., 2^{31} = **2147483648** code positions.
Schemes	**ASCII** and **ISO 646** are examples of 7-bit encoding. In fact, only **English**, **Latin**, and **Swahili** languages can use plain 7-bit ASCII with no additional characters. Most languages based on the Latin alphabet require larger code set.	It covers **most common European languages**, like French or German, that have accented letters, as well as **Arabic** and **Hebrew**. Many national variants were developed. To normalize the mess of 8-bit encodings, ISO came up with the **ISO 8859** series of standards.	It is needed for **Asian languages**, such as Chinese and Japanese that use ideographs, or hieroglyphs, instead of letters. **Windows NT**, for example, uses **16-bit internally** for all character values.	This system, also known as Universal Multiple-Octet Coded Character Set (**UCS**), was developed as standard in 1993. Today, most PCs have 32-bit registers.

Register sizes are rapidly growing to 64 bits. Special codes are now written for the 64-bit chip used in Windows XP.

Encoding schemes

As a Webmaster, you must pay particular attention to the **encoding scheme** (also known as **character set**) that **you use**; each scheme represents characters used in a **different language**.

Encoding schemes

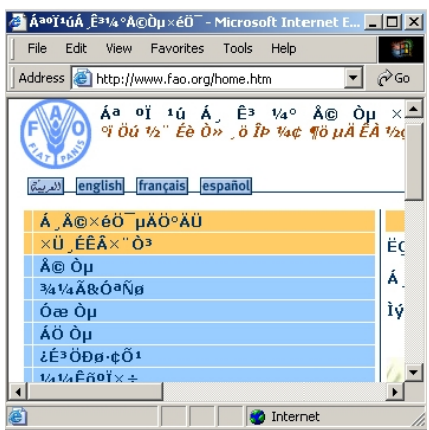


It is recommended that you label a document in the language that it is written in by using the **charset code** in all pages: this will allow browsers to **automatically choose the correct character type** to display, independently from the workstation setting.

In this example, the Chinese version of the FAO Home Page contains a charset code for Chinese and thus this page is automatically displayed in Chinese.

Encoding schemes

If **you don't use** charset code, your result will be like this:



This is the same Chinese web page seen in the previous example.

In order to see this page in the correct font, the user needs to change the document encoding, by selecting **Encoding** from the View Menu and by clicking on **Chinese Simplified**.

This is not the best way to present your information!

Encoding schemes

Using the charset code, you can **insert**, **edit** or **update** text in an HTML page **in the original language** of that page.

The charset code must be included in your HTML page by inserting the following **META tags**:

Employment
Emploi
Empleo

English, French, Spanish:
<meta http-equiv="Content-Type" content="text/html;**charset=ISO-8859-1**">

就业机会

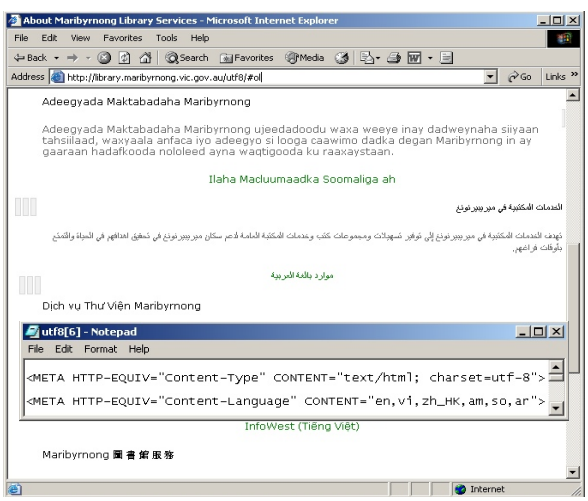
Chinese:
<meta http-equiv="Content-Type" content="text/html; **charset=gb2312**">

فرص العمل

Arabic:
<meta http-equiv="Content-Type" content="text/html;**charset=windows-1256**">

These META tags must be inserted in the **HEAD** section **within the <HEAD> and </HEAD>** sections of the HTML page.

Encoding schemes



You can also use UTF-8 (Unicode Transformation Format, 8-bit encoding form) as encoding, especially when you have to mix languages freely on the same page.

UTF-8 is an encoding form of the Unicode Standard, the universal character encoding standard used for representation of text for computer processing.

The Unicode Standard provides the capacity to encode **all of the characters used** for the written languages of the world.

One disadvantage of Unicode is that it takes more space to store plain text and transmission of Unicode data can therefore use more bandwidth.

Encoding schemes

More information about Unicode

Unicode is now widely used and has become the preferred character set for the Internet, especially for developing, processing and exchanging multilingual HTML and XML documents, and it is also being adopted for use in e-mail.

However, Unicode is not the most common character set in use. According to Microsoft, there is "built-in" support for Unicode in Windows NT and Windows 2000, but only limited support in Windows 95 and Windows 98. In addition, older office automation software such as Office 97, do not offer Unicode options.

Encoding schemes

For pages written, for example, in Arabic, the **direction in which the text** is to be displayed must be specified.

Arabic encoding can appear as follows:

```
<HTML dir="RTL" lang="ar">
<HEAD>
<meta http-equiv="Content-Type"
content="text/html; charset=windows-1256">
</HEAD>
```

RTL means **from right to left**.

If you have a mixed language page, you will need to use **spans** to enclose the Arabic content.
Using span tags (<SPAN>, </SPAN>), you can separate the document into different paragraphs, and apply the RTL only to the Arabic parts.

Guidelines and procedures

In order to maintain original characters when converting **Arabic** and **Chinese Word documents into HTML** on **non-Arabic and non-Chinese workstations**, certain procedures need to be followed.

Descriptions of these procedures can be downloaded and printed below *(note: procedures based on the usage of charset=windows-1256 for Arabic and charset=gb2312 for Chinese):*

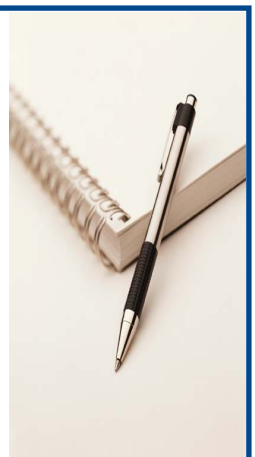
Converting an Arabic document into HTML on a non-Arabic workstation.

Converting a Chinese document into HTML on a non-Chinese workstation.

Procedures similar to the ones described above for Arabic and Chinese documents, can be applied to other documents written in **non-Latin** characters (e.g. Russian documents). The appropriate charsets need to be inserted.

Summary

- In computers, **characters** are **stored** in memory **as numbers**.
- Characters can be **coded in different ways** (**encoding schemes**).
- As a Webmaster, you must **specify which encoding scheme** you are using in order to correctly display the text of your document on the Web.
- You must pay particular attention when converting **Arabic** and **Chinese Word documents into HTML** on **non-Arabic and non-Chinese workstations**.



Exercises

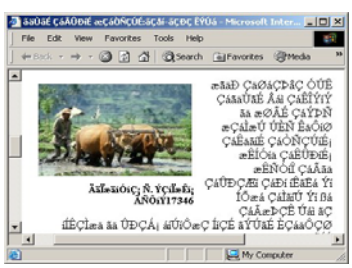
The following three exercises will test your understanding of the concepts covered in the lesson and provide you with feedback.

Good luck!

Exercise 1

You created a web page in Arabic using a charset code. However, your page is being displayed on an English language workstation.

How will your page look?

○ 

○ 

Click on your answer

Exercise 2

What is the function of a charset code?

- ○ It allows you to translate text from an HTML page into a specific language.
- ○ It allows browsers to automatically select the correct character type to display.
- ○ It allows the user to select the correct character type in which to display the Web page.

Click on your answer

Exercise 3

Which of the following examples of HTML script is correct for a Web page written in Arabic?

○
```
<HTML dir="RTL"
lang="ar">
<HEAD>
<meta http-
equiv="Content-Type"
content="text/html;
charset=ISO-8859-1">
</HEAD>
```

○
```
<HTML dir="LTR"
lang="ar">
<HEAD>
<meta http-
equiv="Content-Type"
content="text/html;
charset=windows-
1256">
</HEAD>
```

○
```
<HTML dir="RTL"
lang="ar">
<HEAD>
<meta http-
equiv="Content-Type"
content="text/html;
charset=windows-
1256">
</HEAD>
```

Click on your answer

If you want to know more...

Download and print documents for more information on:

ASCII, ISO 8859-1, Unicode and ISO 10646

Windows and code pages

XML and E-mail encoding

American National Standards Institute (ANSI) http://www.ansi.org/
HTML4-“ HTML 4.01 Specification” http://www.w3.org/TR/REC-html40/
HTTP1.1- “RFC2068 Hypertext Transfer Protocol—HTTP/1.1”
http://www.cis.ohio-state.edu/htbin/rfc/rfc2068.html
International Organization for Standardization (ISO) http://www.iso.org/
International Electrotechnical Commission (IEC) http://www.iec.ch/
Unicode Consortium http://www.unicode.org/
Unicode Technical Report #17: Character Encoding Model.
http://www.unicode.org/unicode/reports/tr17/
Windows Character Set
http://www.microsoft.com/globaldev/reference/sbcs/1252.htm
World Wide Web Consortium (W3C) http://www.w3.org/
XML- “Extensible Markup Language (XML) 1.0 (Second Edition)”-
http://www.w3.org/TR/2000/REC-xml-20001006