B211 Internet Computing

Internet Access for People with Different Cultural Backgrounds

Lecture Outline

- The need for Internationalization
- Character Encoding Standards
- Language attributes in HTML
- Multi-lingual Web Sites
- The International Web

Internationalization (i18n)

- Internationalization is the issue of making the Internet accessible to people from different countries and with different cultural backgrounds.
- The principle technical issue dealing with this is the ability to represent and process information *in different languages*.

The Need for Internationalization

- Historically, most Internet and web sites have information presented in English.
- For the Internet to be truly universal, it needs to be usable by people who do not understand English.
- Even if the majority of people learn rudimentary English, a lot of culturally specific information cannot be converted to English.
    - Eg. literature, names, colloquial terms, etc.

The Need for Internationalization (cont'd)

- In the development of the languages used on the Internet as a whole, it is a case of balancing between:
  - the need to have a universal language most can use, so that there is one less barrier in communications between everyone (*globalization*),

    and
  - the need to have local languages preserved and proliferated, so that there is diversity in information content and presentation (*localization*).

Who's Who in Internationalization

- W3C working groups in its User Interface domain (www.w3.org/International/)
  - specifications for internationalization features in HTML, HTTP, CSS, etc.
- The Unicode Consortium (www.unicode.org)
  - defined the unicode standard.
- United Nations' Universal Networking Language Programme (http://www.unl.ias.unu.edu/)
  - developing a software system consisting the Enconverter / DeConverter, dictionary system, language servers, and procedures.
- ISO (International Standards Organization)
  - certain aspects of character encoding standards

Character Encoding Standards

- For computers to support different languages, there must be ways for characters in the languages to be represented in software.
- Standards in character encoding allow documents containing (printable or non-printable) these characters to be processed consistently by all software.
- The representation at the machine level will be in binary, but we usually look at those binary numbers in their decimal or hexadecimal (base 16) representation.

Character Encoding Standards

- A character encoding standard defines a table where each character is assigned a numeric value – each entry in the table is called a *code point*.
- Eg character encoding standards: ASCII, Unicode, UTF, MIME.

7-Bit ASCII

- Defined in ISO 646 - A widely used representation for plain text.
  - Stands for the American Standard Code for Information Interchange.
  - Also referred to as US-ASCII.
- Strictly speaking, does use 8-bits (one full normal byte), but with the top bit set to 0.
- With 7-bits, the table defines conversions for 128 characters. It encodes all printable characters.

Example 7-Bit ASCII Encodings

Character	Decimal Encoding
…	...
>	62
?	63
@	64
A	65
B	66
C	67
D	68
E	69
…	…

8-Bit ASCII

- 7-Bit ASCII only able to support English type characters – inadequate for other human languages.
- ISO defined another series of ASCII standards called the 8859 series
  - uses 8 bits instead of 7.
  - The first 128 character conversions (with the top bit set to 0) are identical to 7-bit ASCII.
  - It also has an extra 128 conversions.
  - What the extra conversions are depend on the character set.

Languages Covered by ISO 8859

Character Set	Encoding for characters in:
8859-1	Afrikaans, Albanian, Basque, Catalan, Danish, Dutch, English, Faroese, Finnish, French, Galician, German, Icelandic, Irish, Italian
8859-2	Croatian, Czech, Hungarian
8859-3	Esperanto, Maltese
8859-4	Bulgarian, Byellorussian, Macedonian
8859-5	Arabic
8859-6	Greek
8859-8	Hebrew
8859-9	Turkish
8859-10	Lapp, Latvian, Lithuanian

Languages Covered by ISO 8859

- Unfortunately, only one set in the 8859 series (ISO 8859-1) was followed consistently.
  - Most Web browsers support ISO 8859-1 (also called ISO Latin 1) by default.
- Unicode have taken over as the preferred encoding for non-English characters.

Unicode

- Published by the Unicode Consortium.
- Encodes in 16-bits, so can represent 65 536 characters.
- The first 128 code points are identical to 7-bit ASCII.
- The first 256 code points are identical to ISO 8859-1.

Goals of Unicode

- **Universal**: to include most computer and spoken languages. Includes languages not supported in ISO 8859 such as Latin, Chinese, Japanese, Sanskrit, etc.
- **Efficient**: software should be able to process the encoding with minimal processing.
- **Unambiguous**: The same numeric 16-bit value must always give the same encoded character – unlike the ISO 8859 series where it depends on which set is used.

Universal Unicode Encoding

- Unlike ISO 646 (7-bit ASCII) which is geared towards computer representation, Unicode makes certain concessions important for natural language.
- Eg.
  - Unicode makes a distinction between representing minus and hyphen, even though in most keyboards and screens, they appear the same.
  - Unicode can also handle characters that do not flow left-to-right.
  - It provides a definition for Chinese/Japanese/Korean (CJK) ideographs or "characters".

ISO 10646

- Also called the Universal Multiple-Octet Coded Character Set (abbreviated UCS).
- A 32-bit encoding
  - over 2 billion characters possible
  - capable of encoding all possible languages that have ever existed on Earth
- An extension of Unicode
  - first 65 536 characters in ISO 10646 is specified by Unicode

ISO 10646 (cont'd)

- To cater for the inevitable stage where Unicode runs out of code points.

UTF-7 and UTF-8

- UTF stands for Universal Transformation Format
- Some current network hardware and software can only handle 7 or 8-bit transmissions, and therefore hard to transmit Unicode's 16-bit encodings.
- UTF-7 and UTF-8 are a conversion scheme that transforms Unicode to ASCII.
  - The Unicode characters are converted to a series of bytes.
- UTF transformations are reversible.

MIME Character Encodings

- MIME (Multipurpose Internet Mail Extensions) is primarily an email transmission protocol
  - We will go into details of MIME when we talk about how email works in the topic 5.
- MIME it also includes an encoding scheme for characters.
- Besides email, HTTP also uses the MIME encoding for its character representation.

MIME Character Encodings (cont'd)

- MIME has two ways of encoding non-ASCII text into 7-bit ASCII text:
  1. Quoted Printable
  2. Base64
- The principal reason for having MIME was to allow non-ASCII text to be transmitted in emails.
- There are also other methods of transmitting MIME messages: Binary, 7-bit, 8-bit, X-Token
  - more on these when we come to talking about email

MIME Character Encodings (cont'd)

- An example email from my mail box:

```
...
To: h.hiew@central.murdoch.edu.au
Subject: Access to Lecture Overheads
...
MIME-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
...
```

- An example HTTP request with language specifications:

```
GET /project/index.html HTTP/1.1
MIME-Version: 1.0
Content-Language: en-US
Content-Type: text/plain; charset=us-ascii
...
```

US English

MIME Quoted Printable

- For character set which are mostly 7-bits already.
- Leave all characters which have high-bit of 0 alone
- Transforms characters needing a high-bit of 1 to sequence of 2 characters, the first character being "=".

MIME Base64

- Converts each group of 3 characters (24 bits) to 4 ASCII characters (32 bits).
- Renders data unreadable to a human reader.

Putting language attributes in HTML

- The previous discussions about character encodings are about creating a standard representation for software (eg web browsers) and hardware (eg. keyboards) so that we can input/output/process multi-lingual data.
- At a higher level, we can also label HTML documents with the language they are written in.

Putting language attributes in HTML

- Why do it?
  - advanced browsers can use the information.
  - more and more web applications are expected to use the information (eg search engines that will search for documents in a particular language).

Language Attributes in HTML

- Using the language content META tag to label the whole document . Eg. a document in English:

```
<META HTTP-EQUIV=Content-Language CONTENT=en>
```

- Using the LANG attribute for different sections of the document. Eg. A French paragraph:

```
<P lang=“fr”>Ce paragraphe est en Francais. </P>
```


Language Attributes in HTML (cont’d)

- Using the LANG attribute to point to another version of the document.

```
<LINK REL=alternate HREF=mydoc-fr.html LANG=fr
TITLE="La vie souterrainne">

<LINK REL=alternate HREF=mydoc-de.html LANG=de
TITLE="Leben unter Grund">
```

- Using presentation labels. Eg. to mark right-to-left writing direction:

```
<!ENTITY rlm ...>
```

Translating a Web Site to Another Language

- Potential benefits
  - Widening the audience to include speakers of multiple languages if your organization has an international presence.
  - Present an authentic image to foreign audience who may already be familiar with what you offer.
  - Low cost way of testing foreign response to products.
- Potential difficulties
  - Hiring a professional translator who is familiar not in with the source and target languages of the translation, but also with the technical contents of the web site.

Multi-lingual Web Sites

- Different versions of the same web-site in different languages can be linked together by
  - a splash page (a front page) where users get to select their viewing options, in this case the language to view in.
  - a navigation bar allowing users to move to different versions.
- Having multiple versions of the same information means having to update all versions if the information changes.
  - This can be hard unless you always have a translator available for all the different languages.

Multi-lingual Web Sites (cont'd)

- It may be better to maintain a multi-lingual "core" section which is relatively stable, and have the constantly changing sections in one language only.

Multi-Lingual Support in Web Browsers

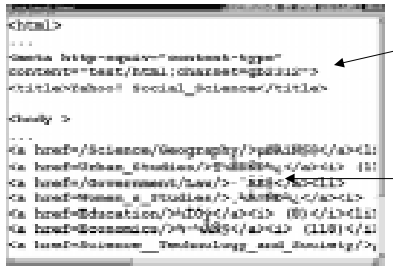
- Browsers translate the HTTP message based on the `charset` parameter in the `Content-Type` field.
  - This field is set by the web browser which serves the pages
- They also looks at the `<META HTTP-EQUIV=Content-Language CONTENT=...>` tag in a HTML document
  - this tag is written by the author of the HTML document
  - remember HTML documents are sent as a HTTP messages
- It depends on the implementation of browser which one it gives more preference to when translating.

Multi-Lingual Support in Web Browsers (cont'd)

- One of the key issues in the multi-lingual support of a web browser is to have the proper fonts for the character sets installed in the operating system.
  - This can be done by prior to accessing the web pages, or
  - using *embedded fonts*, where the fonts required to be installed are sent with the HTML page to the browser, and the browser installs it automatically.

Example of accessing a non-English web site

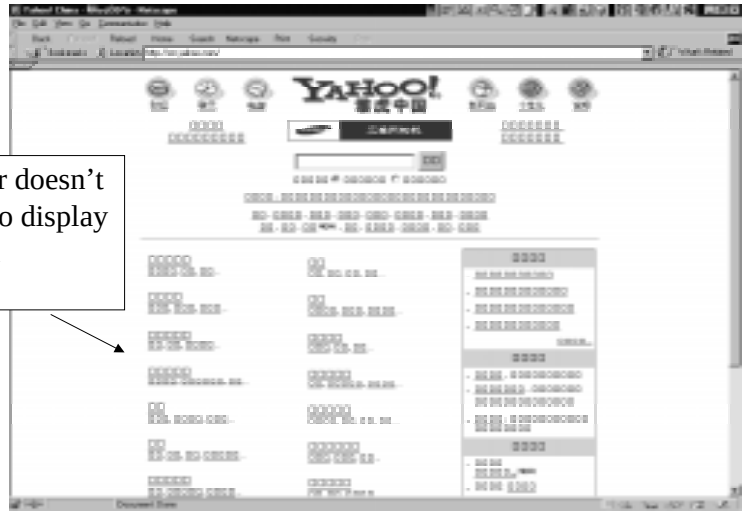
- An example Chinese page in Unicode (cn.yahoo.com).
- When viewed with an ASCII text editor, some characters cannot be displayed properly:



Unicode charset "gb2312"

Non-ASCII characters, incorrectly displayed by the text editor

Eg. in a browser without the appropriate language support



The browser doesn't know how to display the encoded characters

Eg. in a browser with the appropriate language support

Multi-Lingual Support in Web Browsers (cont'd)

- We will look at setting browser options for languages in week 10's lab.