Chapter IR:IX

IX. Acquisition

- ❑ Crawling the Web
- ❑ Conversion
- ❑ Storing Documents

Conversion

File formats

- Text stored in hundreds of incompatible file formats
  - Raw text, RTF, HTML, XML, Microsoft Word, ODF, PDF
- Other types of files also important
  - PowerPoint, Excel
- Typically use a conversion tool
  - Converts the document content into a tagged text format such as HTML
  - Retains some of the important formatting information that would be lost in plain text (words in headings, bold text etc. are important for weighting)
  - You can see this by clicking on the "cached" version of for instance PDF documents on any major search engine
  - For some document types (e.g., PowerPoint) the cached version might look unreadable
  - Still the content is important for indexing, not readability
  - HTML also has the advantage that the user does not need a specific application to show content

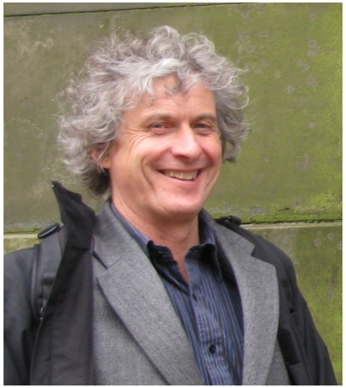
Conversion

Character encoding

- Character encoding is a mapping between bits and glyphs
  - Getting from bits in a file to characters on a screen
  - Can be a major source of incompatibility
- English: ASCII
  - Established 1963
  - Encodes 128 letters, numbers, special characters, and control characters in 7 bits, extended with an extra bit for storage in bytes
- Other European (most): Latin-1 (ISO-8859-1)

Conversion

Character encoding: Documents lie!



Even when documents say they are in ASCII or ISO 8859-1, you have to assume that they are lying, because it's extremely common for such documents to be actually encoded in Windows-1252.

If you assume that characters in the range 128–159 (decimal) are control characters rather than Windows punctuation (smart quotes, em dashes etc.) then your search results will look very messy.

[David Hawking]

Conversion

Character encoding: Even more problems

- Other languages can have many more glyphs
  - Chinese has more than 40,000 characters, with over 3,000 in common use
- Many languages have multiple encoding schemes
  - For instance, CJK (Chinese-Japanese-Korean) family of East Asian languages, Hindi, Arabic
  - Must specify encoding
  - Can’t have multiple languages in one file
- Unicode developed to address encoding problems

Conversion

Unicode

- ❑ Single mapping from numbers to glyphs that attempts to include all glyphs in common use in all known languages
- ❑ Multiple languages possible in one file
- ❑ Many ways to translate Unicode numbers to glyphs
  - UTF-8, UTF-16, UTF-32
- ❑ Proliferation of encodings comes from a need for compatibility and to save space
  - UTF-8 uses one byte for English (ASCII), as many as 4 bytes for some traditional Chinese characters
  - Variable length encoding: more difficult to do string operations (count characters or jump to a position)
  - UTF-32 uses 4 bytes for every character
- ❑ Many applications use UTF-32 for internal text encoding (fast random lookup) and UTF-8 for disk storage (less space)

Conversion

Unicode: UTF-8

Decimal	Hexadecimal	Encoding
0–127	0–7F	`0xxxxxxx`
128–2047	80–7FF	`110xxxxx 10xxxxxx`
2048–55295	800–D7FF	`1110xxxx 10xxxxxx 10xxxxxx`
55296–57343	D800–DFFF	Undefined
57344–65535	E000–FFFF	`1110xxxx 10xxxxxx 10xxxxxx`
65536–1114111	10000–10FFFF	`11110xxx 10xxxxxx 10xxxxxx 10xxxxxx`

- Greek letter pi (π) is Unicode symbol number 960
- In binary, 00000011 11000000 (3C0 in hexadecimal)
- Final encoding is **110**01111 **10**000000 (CF80 in hexadecimal)