Unicode

Ina Minei

What we are going to talk about:

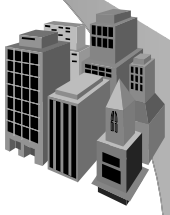
- Before Unicode - some older standards
- Unicode
- After Unicode - ISO 10646

Myths and Reality

- Global village
- Information highway

yes...

If you read
(and write)
English!



Before Unicode...

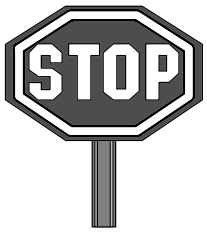
- ASCII (ISO 646)
- ISO 8859 - 12 different codes
- Japanese, Chinese, Korean and Taiwanese multi octet national standards.

ASCII

- American Standard Code for Information Interchange 1967
- The American national version of the ISO 646 standard.
- 7-bit code with range 0-127
- Adopted in 1988 as the ISO 646 standard.

Cons

- English is the only language that can be encoded with ASCII.



Pros

- Simple
- Elegant
- Widely used - English is widely spoken.
- Contains all the special characters used in most programming languages. (e.g. #,@,$) (Actually it is because of ASCII that these symbols are used in programming...)

ISO - 8859

- 8 bit extension of ASCII.
- Backwards compatible with ASCII.
- The most important standard for coding European languages.

Problem:

The European languages use together more than 256 different characters, although each by itself uses less than 256.

Solution:

The languages are divided in 12 groups according to common commercial interests . Each group has a separate standard:
ISO 8859- n (n = 1,2... 12)

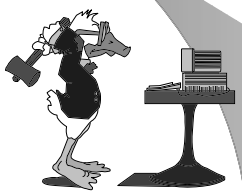
ISO LATIN 1 (ISO 8859-1)

- ISO 8859 standard for Western European countries.
- Encodes: Danish, Dutch, English, Finnish, French, German, Italian, Icelandic, Norwegian, Portuguese, Spanish, Swedish.

Other ISO 8859 standards

- ISO 8859 - 6 - Latin/Arabic
- ISO 8859 - 7 - Latin/Greek
- ISO 8859 - 8 - Latin/Hebrew
- ISO 8859 - 9 (ISOLATIN 5) - a variant of ISOLATIN 1 for Turkish.

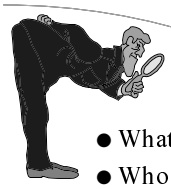
It is difficult to transform documents between the various encoding methods.



The solution:

A single, uniform standard that can encompass all languages.

UNICODE !



Unicode

- What - the goals of Unicode
- Who - the people involved
- How - the basic principles of Unicode
- Special topics

What is Unicode

- A simple uniform standard that can encode all the living languages in the world and that can form a long-lasting basis for standardized world-wide information systems.

Unicode - who

- Industry initiative - The Unicode project was started in 1989 by a working group of professionals from : Xerox, Apple, Microsoft, NeXT, Sun, IBM and others.
- In 1990 the Unicode Consortium was created.
- In 1991 the Consortium was incorporated as a non-profit organization, Unicode Inc.

Unicode - Basic Principles

- All characters are 16 bit wide.
- Full encoding. All the available 65536 codes can be used to represent characters.
- Unicode characters are international and electronic. They do not belong to any natural language.

Basic Principles (continued)

- Characters that are equivalent in form, usage and essential properties are unified across languages.
- Letters and symbols with common shapes and different functions are given separate codes.
- Unicode characters are electronic codes, not glyphs. Glyphs are the visual representation of characters. There are far more glyphs than characters.

Code grouping

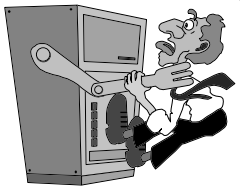
- The first 256 codes follow precisely the arrangement of ASCII and ISO LATIN 1.
- Codes unique to a language are grouped together in standard order.
- Codes that represent letters, punctuation marks and symbols shared by multiple languages are grouped together.
- Asian characters are grouped by frequency of occurrence.

E-mail encoding problems

- Unicode characters are 16 bit wide.
- Internet mail currently only supports 7-bit US ASCII as a character set.
- Unicode is not a permitted character set , and it must be encoded differently for transfer over e-mail.

Encoding - the solution

- The UTF-7 encoding standard encodes Unicode characters as US-ASCII, together with shift sequences to encode characters out of range.



Coverage

- Unicode contains 34168 distinct codes in 24 distinct scripts.
- Unicode does NOT cover 15 modern scripts, among which are : Cambodian, Mongolian and Ethiopic.
- Logos, graphics and private-use characters are systematically excluded.

Language tagging

- Unicode characters are language neutral.
- We need to know the language of the document if we want to do:
  - spell checking
  - hyphenation
  - grammar checking

Language tagging (continued)

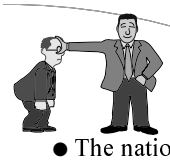
- A higher level protocol must be used to specify the language.
- Several standards for representing names of languages already exist.
  - ISO 639 (1988) - two letter codes of about 140 languages, ISO 639-2 (1991) - three letter extension.
  - ISO 11639 - three letter codes

Language tagging (continued)

- Issues to address:
  - special values for languages. E.g. : machine language, mixture of languages, a human language for which no code exists.
  - means for indicating the range of a language in a multilingual text.
  - means for indicating subdivisions of a language, for example different dialects, different orthographies.

Unicode critics

- The coding spaces reserved for the square characters are set at ~45000. However, the Chinese Character Code for Information Interchange has a character set of 75684 characters.
- The square character set is an open set with an average of 114 characters being added every year.



Critics (continued)

- The national prides of the Far Eastern Countries are not taken into account: the Japanese, Chinese and Korean character sets were unified.
- There is no Far Eastern user participation in the code design or character selection.
- "We are not dogs under the American table" - said a member of the Korean Government Standard Committee...

Critics (continued)

- In the alphabetical languages, Unicode separates characters with different meaning but similar shape. In Asian languages it fails to do the same, possibly due to the ignorance of the code designers.

Critics (continued)

- Each character appears only once, no matter how many alphabets use it. Thus, there is only one A, one blank, one period. For Latin, this does not influence sort.
- In CJK each Chinese character appearing in each of the alphabets appears only once. The ordering of the character is different for Chinese, Japanese and Korean and the unification destroys the order.

Is Unicode used?

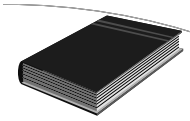
- Windows NT - uses Unicode and provides a font with about 1750 glyphs.
- The Mule editor, a multilingual, 16 bit extension of Gnu emacs - uses Unicode.
- Y & Y Inc. sells a Unicode font.
- Gamma Productions offers utilities such as spell checkers for all the languages coded by Unicode. It also provides fonts and correct ligatures.

Is Unicode used?

- MASS - Multilingual Application Support Service is an environment for developing multilingual applications.
- Unicode can be used as the basic alphabet on the Web. The document must be language tagged to specify its language.

ISO / IEC 10646

- 32 bit code.
- Unicode corresponds to the 16 bit version, where the first two octets are equal to 0.



Bibliography

- The Unicode Standard
- The Unicode homepage www.stonehand.com/unicode.html
- comp.text, soc.culture.china, soc.culture.japan, soc.culture.korean, soc.culture.hongkong, soc.culture.taiwan
- Cahiers GUTenberg - Codage des caracteres, d'ASCII a UNICODE