What is Unicode?
An Overview

TD Core AS&DM I18N Technology (AG)
May 15, 2009
1. What are Code Pages?

2. What is Unicode?

3. Unicode@SAP

4. Thinking Unicode
Contents

1. What are Code Pages?

2. What is Unicode?

3. Unicode@SAP

4. Thinking Unicode
## From Characters to Code Pages

<table>
<thead>
<tr>
<th>Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Character</td>
<td>a, b, c, A, B, ...</td>
</tr>
<tr>
<td>Glyph</td>
<td>a, a, a, ...</td>
</tr>
<tr>
<td>Coded character</td>
<td>A=65, B=66, ...</td>
</tr>
<tr>
<td>Character Set</td>
<td>A set of characters, to be used together (e.g. Latin alphabet)</td>
</tr>
<tr>
<td>Code page</td>
<td>A set of coded characters (e.g. ISO-8859-1, Shift-JIS)</td>
</tr>
<tr>
<td>Locale</td>
<td>Code page + properties and rules (e.g. isdigit, collation, ...)</td>
</tr>
</tbody>
</table>
A single byte code page assigns one distinct byte to each character. The code points range from 0 to 255. Such a code pages is drawn in a 16 * 16 matrix.

'Ö' = x'D6'
= 13 * 16 + 6

'Ö' is the character D6 is the code point
Asian and Other 16-bit Code Pages

They are drawn in a 256 * 256 matrix.

Example:
The double-byte characters of the Shift-JIS standard

First byte is 80..A0 or E0..FF
Second byte is 40..FF

7-Bit area 00..7F identical to ISO 8859-x

Exchange of English text possible
1. What are Code Pages?

2. What is Unicode?

3. Unicode@SAP

4. Thinking Unicode
From Code Pages to Unicode

Every standard code page supports only a certain group of languages (e.g. Western European, Eastern European, Japanese)

Within one computer system only one code page can be supported in a clean way

Therefore a universal code page that supports all letters, punctuation signs, technical symbols etc. of all languages is required
Unicode Principles

Unicode is a superset of all existing character sets

Unicode encodes plain text (no rendering information)

Unicode defines characters, not glyphs (semantics, not visual representation)

Unicode unifies characters used in different scripts (CJK* Unification)

*Chinese, Japanese, Korean
The Unicode ‘Code page’

Space for 1,000,000 characters

64,000 characters coded by one 16bit code point

Further characters coded by two 16bit code points (surrogates)
Unicode Characters

- ASCII
- General Scripts
- Symbols
- CJK Ideographs
- Hangul
- Compatibility
- Surrogate Area

65,000 characters

Additional 1,000,000 characters
Unicode Encoding Forms

**UTF-8**
- byte-based encoding scheme
- one character is coded with 1-4 bytes
- compatible with 7-bit ASCII

**UTF-16**
- 16bit units
- often used characters occupy one 16bit unit
- further characters are coded with two 16bit units

**UTF-32**
- 32bit units
- fixed size for all characters

All encoding forms support the same amount of characters.
## Coding example

<table>
<thead>
<tr>
<th>Character</th>
<th>ISO-8859-1</th>
<th>UTF-8</th>
<th>UTF-16</th>
<th>UTF-32</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0x41</td>
<td>0x41</td>
<td>0x0041</td>
<td>0x00000041</td>
</tr>
<tr>
<td>Ä</td>
<td>0xC4</td>
<td>0xC3 0x84</td>
<td>0x00C4</td>
<td>0x000000C4</td>
</tr>
</tbody>
</table>
Contents

1. What are Code Pages?

2. What is Unicode?

3. Unicode@SAP
UTF-8
- for external communication (e.g., file, network)
  - no endian problems
  - minimum average data size
  - limited backward compatibility to non-Unicode systems

UTF-16
- internal (in memory)
  - best compromise between memory usage and algorithmic complexity
  - fits to Java and Microsoft environment
  - best way to migrate existing ABAP and C programs