Unicode: a tool for system interoperability and human communication

Cathie Jilovsky
What is Unicode?

Unicode provides a unique number for every character, no matter what the platform, no matter what the program, no matter what the language.

• enables the creation of products which perform consistently regardless of the language of the user interface
Development of Unicode

• one unique code is used to represent each character, even if that character is used in multiple languages
• developed for maintenance rationalisation and economic reasons
• provides a platform for multi-language systems
<table>
<thead>
<tr>
<th>Code</th>
<th>Character</th>
<th>Code</th>
<th>Character</th>
<th>Code</th>
<th>Character</th>
<th>Code</th>
<th>Character</th>
</tr>
</thead>
<tbody>
<tr>
<td>0000</td>
<td>NUL</td>
<td>0001</td>
<td>DC1</td>
<td>0002</td>
<td>STX</td>
<td>0003</td>
<td>ETX</td>
</tr>
<tr>
<td></td>
<td>DLE</td>
<td></td>
<td>!</td>
<td></td>
<td>DC2</td>
<td></td>
<td>DC3</td>
</tr>
<tr>
<td></td>
<td>SP</td>
<td></td>
<td>1</td>
<td></td>
<td>&quot;</td>
<td></td>
<td>#</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td></td>
<td>A</td>
<td></td>
<td>2</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>@</td>
<td></td>
<td>Q</td>
<td></td>
<td>B</td>
<td></td>
<td>C</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td></td>
<td>a</td>
<td></td>
<td>R</td>
<td></td>
<td>S</td>
</tr>
<tr>
<td></td>
<td>`</td>
<td></td>
<td>q</td>
<td></td>
<td>b</td>
<td></td>
<td>c</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Unicode and ASCII
### Unicode examples

#### Musical Symbols

<table>
<thead>
<tr>
<th>1D100</th>
<th>1D11</th>
<th>1D12</th>
<th>1D13</th>
<th>1D14</th>
<th>1D15</th>
<th>1D16</th>
<th>1D17</th>
<th>1D18</th>
<th>1D19</th>
<th>1D1A</th>
<th>1D1B</th>
<th>1D1C</th>
<th>1D1D</th>
<th>1D1E</th>
<th>1D1F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- $\sin \alpha = \sin \beta = \sin \gamma$.  (Law of Sines).
- $\cos \alpha = \cos \beta \cos \gamma + \sin \beta \sin \gamma \cos \alpha$.  (Law of Cosines)
- $\tan \frac{\alpha}{2} = \sqrt{\frac{\sin (s - b) \sin (s - c)}{\sin s \sin (s - c)}}$, where $s = \frac{1}{2}(a + b + c)$.  

---

**CAVAL Collaborative Solutions**
Using OSCAR (Unicode)

OSCAR (Unicode) looks exactly like the regular OSCAR library catalog except that it will allow you to display and search in Arabic, Chinese, Japanese and Korean as long as you are using a PC that is configured to use Unicode. For more information on Unicode, see [Unicode Consortium](http://www.unicode.org).

**Using OSCAR (Unicode) in the Main Library:**

Use one of the computers on the third floor (Room 320) of the Main Library.

Note: Make sure Internet Explorer is set to display Unicode by clicking on “View” on the menu bar at the top of the window. Click on Encoding. Make sure Auto-select is not selected. Select Unicode (UTF-8).

**Using OSCAR (Unicode) from your home or office computer:**

In order to display and search Arabic, Chinese, Japanese or Korean scripts and characters you must have the following:

1. Unicode support within the operating system
Unicode and MARC

MARC 21 Specifications for Record Structure, Character Sets, and Exchange Media

Library of Congress
Network Development and MARC Standards Office

Web Version
January 2000

Table of Contents

General Record Structure
Character Sets:
   Part 1: MARC-8 Environment
   Part 2: UCS/Unicode Environment
   Part 3: Code Tables and Lists

Unicode and MARC
About CAVAL

• 25 years of history
• Range of services and products
• Range of customers
• Open to further discussions and possibilities
• Looking for Collaborative Solutions
At CAVAL
Multi-languages at CAVAL
Multi-lingual solutions

• Over 60 languages (more than OCLC!)
• Roman and non-Roman scripts
• Translation and transliteration
• Print and electronic resources
• Web sites, documents
• Reference work / Research
• Abstracting
Multi-language environment
Aleph implementation

1. Storage issues
2. Conversion issues
3. Data input issues
4. Display issues
1. Storage issues

- Aleph system stores data in Unicode format
- Conversion needed when importing and exporting records
- Extensive testing with Kinetica
2. Conversion Issues

- KineticaWeb – does not currently support full MARC8 character set
- MARC8 character set does not contain all Unicode characters
- Some display issues – but underlying data values correct
- Developed conversion routines but took time
3. Data input issues

- Tamil vernacular script
- Input/edit in Aleph Client
  
  ![Tamil Text](image)

- Display in Aleph Web OPAC

- Sometimes 2 characters need to be input to achieve correct display
  
  ![Tamil Text](image)
4. Display issues

- Vietnamese example - double diacritics (low mark + high mark)
- Displayed in Aleph Client with underline

24500 |a Sự tích ông táo |h [videorecording]

- Aleph Web OPAC – correct display
Display issues (cont)

- Vietnamese (same record)
- Input/edit in Kinetica Client

245 00 ǂaS.ǂw tích ông tạo ǂh[videorecording].

- Display in Kinetica Web

245 00 ǂa S. ǂw tích ông tạo ǂh [videorecording].
More Display Issues ...

• Thai example - double diacritics (letter + low mark + high mark combinations)

• Aleph Client display

260  a [Bangkok]:
    b Khana ’Anukammakan Fai Khommun lae Nangsu Thiraluk, Khana Kammakand Damnoen
    Ngan Khunchalorang Khrop Rop 100 Phaeng Kansathapan Rongrian Phaet,
    c 2533 [1990].

• Aleph Web OPAC – correct display

260  a [Bangkok]: b Khana ’Anukammakan Fai Khommun lae Nangsu Thiraluk, Khana Kammakand
    Damnoen Ngan Khunchalorang Khrop Rop 100 Phaeng Kansathapan Rongrian Phaet, c 2533
    [1990].
<table>
<thead>
<tr>
<th>FMT</th>
<th>BK</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDR</td>
<td>nam 22</td>
</tr>
<tr>
<td>001</td>
<td>000156893</td>
</tr>
<tr>
<td>005</td>
<td>20020702184228.0</td>
</tr>
<tr>
<td>008</td>
<td>020702s2001 ii 000 0</td>
</tr>
<tr>
<td>035</td>
<td></td>
</tr>
<tr>
<td>040</td>
<td></td>
</tr>
<tr>
<td>1001</td>
<td></td>
</tr>
<tr>
<td>24510</td>
<td></td>
</tr>
<tr>
<td>260</td>
<td></td>
</tr>
<tr>
<td>300</td>
<td></td>
</tr>
<tr>
<td>650.0</td>
<td></td>
</tr>
</tbody>
</table>
Further development

• incorporation of native vernacular scripts
• importing of records from other Unicode systems
• exporting of records to other Unicode systems
• recommend phased approach
CAVAL Collaborative Solutions

www.caval.edu.au

CARM OPAC – carm.caval.edu.au