Unicode and the Web

Nathan Schneider
In our interactions with computers, it is often desirable to use characters other than the standard English alphabet and common punctuation.

When do we use different forms notation?
- Other languages with slightly or completely different alphabets
- Mathematical and scientific notation (e.g. chemical compounds)
Special Text

– Notation particular to a specific field
– Graphical features, such as arrows and bullets, that help us organize information

• Sometimes it’s appropriate to use graphics or special software to view and edit this text, but ideally it should be fairly easy to put special text onto a web page so that it displays correctly and can be edited, copied/pasted, displayed in different sizes and styles, and laid out properly without special software
Ways to Enter Text

- Directly associate keys on the keyboard with characters
- Use a sequence of keys (e.g. Ctrl+'+e => é)
- Represent it with other characters already on the keyboard (e.g. transliterating Egyptian Arabic with Latin characters)
Ways to Enter Text

• Use some graphical mechanism or special software to select characters (e.g. Windows Character Map)
• Scan it from some printed or digital format (e.g. Optical Character Recognition)
• Write it with a stylus: Handwriting Recognition
• Voice recognition technology
Ways to Enter Text

• Each of these methods has advantages and disadvantages
  – Scanning, handwriting, and voice recognition may be easier to use (more natural) but less reliable technologies, ESPECIALLY for “non-standard” text
  – Typing and graphical character selection may be cumbersome and time-consuming
Goal

• In order to ensure that computers will make our lives easier, in part by simplifying and enhancing our ability to communicate, we need to overcome these obstacles

• Computers need to (1) support special text and display it properly, as well as (2) provide convenient and reliable mechanisms for us to input special text
Old Implementation of Text

- Limitations of older computers/software: support of special text
  - Originally, most computers only supported what is known as the ASCII character set. (American Standard Code for Information Interchange) ASCII-I contains 128 characters: some control characters, and all the letters and punctuation that appear on standard American keyboards.
  - Computers see each character as a number. Capital A, for example, is 65. A space is 32. ASCII contains a newline character, a tab character, and (oddly enough) a “beep” character.
Old Implementation of Text

– ASCII-II, ANSI (American National Standards Institute), and other character sets came about later

– This was sufficient for writing computer programs, but not designed for personal use by people around the world
Problems

• There were many problems with attempts to use characters beyond the standard ASCII characters on American keyboards
  – Different computer and software systems used different representations for characters, making it difficult to translate between them
  – Using special fonts to display certain characters (where the computer sees A-Z, etc. but the font displays them as something else) restricts users to a particular font
ISO-8859

- ISO-8859: This is a group of character sets established by the International Standards Organization which implements various languages by mapping several sets of characters to a single range of numeric values. This leaves it up to the viewer (i.e. a web browser) to determine which set of characters to display.

- Latin1 (West European); Latin2 (East European); Latin3 (South European); Latin4 (North European); Cyrillic; Arabic; Greek; Hebrew; Latin5 (Turkish); Latin6 (Nordic)

See http://wwwwbs.cs.tu-berlin.de/user/czyborra/charsets/
It’s all online at http://www.unicode.org
Unicode

• In order to solve this problem, experts have worked over the past 10 or so years to develop what’s known as The Unicode Standard. This seeks to standardize how the computer recognizes special characters. The most recent version is 4.0.
  – It does this by creating a unique identifier (a hexadecimal number) for each character in the system

It’s all online at http://www.unicode.org
The New Way

• Unicode maps only ONE character to each numeric value, which requires more memory (if a large number of characters are to be supported), but makes things MUCH less confusing
  – Hey, memory is cheap now anyway
• It is standardized so that it should be consistent regardless of the user’s platform or system configuration

It’s all online at http://www.unicode.org
Code Points

• A Unicode code point is a hexadecimal number identifying a particular character
  – Hexadecimal is a base-16 system (as opposed to binary, or the base-10 decimal system that we normally use); hexadecimal numbers are sometimes prefixed with 0x or x
  – In hexadecimal (“hex”), the letters A – F represent the values 10 – 15
  – $0x215C = 12+5(16)+1(16^2)+2(16^3) = 8540$
  – $16^4-1 = 655535$ possibilities (with 4 digits)
<table>
<thead>
<tr>
<th>Code</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>05A2</td>
<td>&lt;reserved&gt;</td>
<td></td>
</tr>
<tr>
<td>05A3</td>
<td>HEBREWC ACCENT MUNAH</td>
<td></td>
</tr>
<tr>
<td>05A4</td>
<td>HEBREWC ACCENT MAHAPAKH</td>
<td></td>
</tr>
<tr>
<td>05A5</td>
<td>HEBREWC ACCENT MERKHA</td>
<td>= yored</td>
</tr>
<tr>
<td>05A6</td>
<td>HEBREWC ACCENT MERKHA KEFULA</td>
<td></td>
</tr>
<tr>
<td>05A7</td>
<td>HEBREWC ACCENT DARGA</td>
<td></td>
</tr>
<tr>
<td>05A8</td>
<td>HEBREWC ACCENT QADMA</td>
<td>= azla</td>
</tr>
<tr>
<td>05A9</td>
<td>HEBREWC ACCENT TELISHA QETANA</td>
<td></td>
</tr>
<tr>
<td>05AA</td>
<td>HEBREWC ACCENT YERAH BEN YOMO</td>
<td>= galgal</td>
</tr>
<tr>
<td>05AB</td>
<td>HEBREWC ACCENT OLE</td>
<td></td>
</tr>
<tr>
<td>05AC</td>
<td>HEBREWC ACCENT ILUY</td>
<td></td>
</tr>
<tr>
<td>05AD</td>
<td>HEBREWC ACCENT DEHI</td>
<td></td>
</tr>
<tr>
<td>05AE</td>
<td>HEBREWC ACCENT ZINOR</td>
<td>= tsinor; zarqa</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• This character is to be used when Zarqa or Tsinor are placed above left.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>→ 0598 ☺ hebrew accent zarqa</td>
</tr>
<tr>
<td>05AF</td>
<td>HEBREWC MARK MASORA CIRCLE</td>
<td></td>
</tr>
<tr>
<td>05B0</td>
<td>HEBREWC POINT SHEVA</td>
<td></td>
</tr>
<tr>
<td>05B1</td>
<td>HEBREWC POINT HATAF SEGOL</td>
<td></td>
</tr>
<tr>
<td>05B2</td>
<td>HEBREWC POINT HATAF PATAH</td>
<td></td>
</tr>
<tr>
<td>05B3</td>
<td>HEBREWC POINT HATAF QAMATS</td>
<td></td>
</tr>
<tr>
<td>05B4</td>
<td>HEBREWC POINT HIRIQ</td>
<td></td>
</tr>
<tr>
<td>05B5</td>
<td>HEBREWC POINT TSERE</td>
<td></td>
</tr>
<tr>
<td>05B6</td>
<td>HEBREWC POINT SEGOL</td>
<td></td>
</tr>
<tr>
<td>05B7</td>
<td>HEBREWC POINT PATAH</td>
<td>= zade</td>
</tr>
<tr>
<td>05C4</td>
<td>HEBREWC MARK UPPER DOT</td>
<td></td>
</tr>
<tr>
<td>05D0</td>
<td>☺ HEBREWC LETTER ALEF</td>
<td>= aleph</td>
</tr>
<tr>
<td></td>
<td></td>
<td>→ 2135 ☺ alef symbol</td>
</tr>
<tr>
<td>05D1</td>
<td>ב HEBREWC LETTER BET</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>→ 2136 ב bet symbol</td>
</tr>
<tr>
<td>05D2</td>
<td>ג HEBREWC LETTER GIMEL</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>→ 2137 ג gimel symbol</td>
</tr>
<tr>
<td>05D3</td>
<td>ד HEBREWC LETTER DALET</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>→ 2138 ד dalet symbol</td>
</tr>
<tr>
<td>05D4</td>
<td>ה HEBREWC LETTER HE</td>
<td></td>
</tr>
<tr>
<td>05D5</td>
<td>ו HEBREWC LETTER VAV</td>
<td></td>
</tr>
<tr>
<td>05D6</td>
<td>ז HEBREWC LETTER ZAYIN</td>
<td></td>
</tr>
<tr>
<td>05D7</td>
<td>ח HEBREWC LETTER HET</td>
<td></td>
</tr>
<tr>
<td>05D8</td>
<td>ט HEBREWC LETTER TET</td>
<td></td>
</tr>
<tr>
<td>05D9</td>
<td>י HEBREWC LETTER YOD</td>
<td></td>
</tr>
<tr>
<td>05DA</td>
<td>ק HEBREWC LETTER FINAL KAF</td>
<td></td>
</tr>
<tr>
<td>05DB</td>
<td>ה HEBREWC LETTER KAF</td>
<td></td>
</tr>
<tr>
<td>05DC</td>
<td>ל HEBREWC LETTER LAMED</td>
<td></td>
</tr>
<tr>
<td>05DD</td>
<td>מ HEBREWC LETTER FINAL MEM</td>
<td></td>
</tr>
<tr>
<td>05DE</td>
<td>נ HEBREWC LETTER MEM</td>
<td></td>
</tr>
<tr>
<td>05DF</td>
<td>ס HEBREWC LETTER FINAL NUN</td>
<td></td>
</tr>
<tr>
<td>05E0</td>
<td>ה HEBREWC LETTER NUN</td>
<td></td>
</tr>
<tr>
<td>05E1</td>
<td>י HEBREWC LETTER SAMEKH</td>
<td></td>
</tr>
<tr>
<td>05E2</td>
<td>ע HEBREWC LETTER AYIN</td>
<td></td>
</tr>
<tr>
<td>05E3</td>
<td>צ HEBREWC LETTER FINAL PE</td>
<td></td>
</tr>
<tr>
<td>05E4</td>
<td>כ HEBREWC LETTER PE</td>
<td></td>
</tr>
<tr>
<td>05E5</td>
<td>ר HEBREWC LETTER FINAL TSADI</td>
<td></td>
</tr>
</tbody>
</table>

**Based on ISO 8859-8**

**Points and punctuation**

- HEBREW POINT SHEVA
- HEBREW POINT HATAF SEGOL
- HEBREW POINT HATAF PATAH
- HEBREW POINT HATAF QAMATS
- HEBREW POINT HIRIQ
- HEBREW POINT TSERE
- HEBREW POINT SEGOL
- HEBREW POINT PATAH
  - furtive patah is not a distinct character
00B0 ° DEGREE SIGN
   • this is a spacing character
   → 02DA ° ring above
   → 030A ° combining ring above
   → 2070 ° superscript zero
   → 2218 ° ring operator
00B1 ± PLUS-MINUS SIGN
   → 2213 ± minus-or-plus sign
00B2 ² SUPERSCRIPT TWO
   = squared
   • other superscript digit characters:
     2070 ² –2079 ³
     → 00B9 ¹ superscript one
     ≈ <super> 0032 2
00B3 ³ SUPERSCRIPT THREE
   = cubed
   → 00B9 ¹ superscript one
   ≈ <super> 0033 3
00B4 ´ ACUTE ACCENT
   • this is a spacing character
   → 02B9 ´ modifier letter prime
   → 02CA ´ modifier letter acute accent
   → 0301 ´ combining acute accent
   → 2032 ´ prime
   ≈ 0020 ´ 0301 ´
00B5 µ MICRO SIGN
   ≈ 03BC µ greek small letter mu
00B6 ¶ PILCROW SIGN
   = PARAGRAPH SIGN
   • section sign in some European usage
   → 204B ¶ reversed pilcrow sign
   → 2761 ¶ curved stem paragraph sign ornament
00BB » RIGHT-POINTING DOUBLE ANGLE QUOTATION MARK
   = RIGHT POINTING GUILLEMET
   • usually closing, sometimes opening
   → 226B » much greater-than
   → 300B » right double angle bracket
00BC ¼ VULGAR FRACTION ONE QUARTER
   • bar may be horizontal or slanted
   • other fraction characters: 2153 ½ –215E ¾
   ≈ 0031 ¼ 2044/ 0034 4
00BD ½ VULGAR FRACTION ONE HALF
   • bar may be horizontal or slanted
   ≈ 0031 ½ 2044/ 0032 2
00BE ¾ VULGAR FRACTION THREE QUARTERS
   • bar may be horizontal or slanted
   ≈ 0033 ¾ 2044/ 0034 4
00BF ¿ INVERTED QUESTION MARK
   = turned question mark
   • Spanish
   → 003F ¿ question mark

Letters
00C0 À LATIN CAPITAL LETTER A WITH GRAVE
   ≈ 0041 À 0300 ´
00C1 Â LATIN CAPITAL LETTER A WITH ACUTE
   ≈ 0041 Â 0301 ´
00C2 Å LATIN CAPITAL LETTER A WITH CIRCUMFLEX
   ≈ 0041 Å 0302 ´
00C3 Ä LATIN CAPITAL LETTER A WITH TILDE
Browse Unicode Character Charts

- [http://www.unicode.org/charts](http://www.unicode.org/charts)
How the Web Works

- You type in the URL of the site you want (or click on a hyperlink)
- Your browser requests the IP address of the site with that DNS name
- Your browser sends a page request to the server
- The server generates the page (perhaps a script) and your computer downloads it
- Your browser displays the page
HTML in a Nutshell

- HTML is the standard language that browsers read to display web pages
- It stands for **Hypertext Markup Language**
- Consists primarily of tags surrounding text
  - `<b>my text goes here</b>` - bold
  - `Line1<br>Line2 blah <br>Line 3`
- CSS (**Cascading Style Sheets**) – often used to “style” the text (fonts, colors, positioning, etc.)

➢ Click on “View Source” in your browser
Using Unicode on Web Pages

• Fortunately, HTML offers us a convenient way to represent special characters on web pages

• *HTML Entities* begin with an ampersand (&) and end with a semicolon; there are built-in *named entities*, and designers can specify Unicode characters by entering the character’s number after the # sign
Sample HTML Entities

- The five most important entities essentially “escape” the characters that have significance in HTML:
  - `&lt;` (less than) displays as `<
  - `&gt;` (greater than) displays as `>`
  - `&amp;` displays as `&`
  - `&quot;` displays as "
  - `&apos;` displays as ' (for XML/XHTML only)
Sample HTML Entities

• Others include: `&infin;`, `(8)`, `&hellip;`, (horizontal ellipsis, ...), `&copy;`, `(©)`, `&aacute;`, (à), `&Euml;`, (È), `&ucirc;`, (û), `&Ccedil;`, (Ç), `&ntilde;`, (ñ)
  – Note that some of these are case-sensitive

• Numbered entities for Unicode: `&#8359;` or `&#x20A7;` (Peseta), `&#x069C;` –?

• One drawback is that each font only supports a limited number of characters
  – Arial Unicode MS has broad Unicode support
Demo

• My encoder tool
  – What it does
  – Which encodings it supports
  – Symbols, X-SAMPA example
  – ISO-8859 example
  – Hebrew example
  – Written using: PHP (server-side), HTML/CSS/Javascript (client-side)
  – Show the code
  – Show the dictionary files