PLA Tech Note: Unicode

Unicode: From Chinese to Cherokee; from Kana to Klingon

Most librarians are now familiar with the concept of ASCII code. Mindful of the famed bumper sticker ("Relax, it's only ones and zeroes"), we can describe ASCII as representing each letter, number, or special character by a seven- or eight-bit number that enables computers to talk with one another. ASCII is in many ways *the* killer app: it is the basis for email and was the original basis for HTML. (HTML now uses the Universal Character Set, which defines thousands of characters.)

But what happens when you need to use a different alphabet or set of accents, or even a different script entirely? What happens when the ASCII code for a letter in one language is a different letter in another? There are only 256 possible combinations in ASCII, and some of them map to different characters in different languages. So, about a dozen years ago, Unicode was born, one code to "provide[s] a unique number for every character, no matter what the platform, no matter what the program, no matter what the language."

Unicode is super-ASCII. Instead of a seven-bit binary number, Unicode is 16-bits. If you do the math, this increases the number of descriptive bits many thousandfold.

What is Unicode?
"Unicode ... [defines] a single set of characters for all the world's written languages and [describes] how to use these characters in computer-based writing." (Hoffman, see bibliography). "The Unicode Standard is the universal character encoding standard used for representation of text for computer processing." Unicode "enables a single ... website to be targeted across multiple platforms, languages and countries without re-engineering." The Unicode Consortium's web site, the source of both of these quotes, is a rich source of information on all aspects of Unicode. Founded in 1991, the Unicode Consortium comprises a broad spectrum of corporations and organizations working "at the leading edge of standardizing international character encoding."

Joan Aliprand of the Research Libraries Group (RLG) is one of the authors of *The Unicode Standard, Version 3.0*. She did a presentation in December 1998 for LITA that describes Unicode, its functions and uses, in a straightforward and comprehensible manner, Unicode: Looking Ahead. It is perhaps the clearest and most accessible description available of what Unicode is and does, although it is now slightly dated.

In describing Unicode, Aliprand notes that there is even a place for characters that do not yet have assigned codes, called “Private Use.” She says “There is a proposal [to the Unicode Consortium] for the Klingon script of *Star Trek*, but if you wanted to use Klingon today it would go here [in Private Use].”

Librarians have devoted much professional time to developing transliteration systems for Cyrillic, EACC (East Asian Character Code, for Chinese, Japanese, and Korean), and other ways of fulfilling our ancient mission of cataloging and making accessible. As computers became more and more integral to pretty much everything that libraries do, a way of providing specific and accurate information across languages became crucial, and not only for libraries alone.

Public libraries, as their catalogs expand to include many the character sets of many languages (from Chinese to Cherokee, from Japanese Kana script to *Star Trek*'s Klingon) and as their population's needs become increasingly diverse linguistically, need to think about Unicode as part of the many issues involved in ongoing development of their public access catalogs.

Unicode, EACC, MARC, and Microsoft

It is commerce, of course, not just the needs of libraries, that brought the push for Unicode to the present -- the need for a single software application that could be used worldwide. Microsoft, Apple, and IBM have adopted it (Netscape, Internet Explorer, and Microsoft Office 2000 use it) and XML and Java require it.

Libraries and vendors are working to map MARC records to Unicode, and to translate the traditional library EACC code to Unicode. Public libraries need to work with their public access catalog vendors to get support for displaying, retrieving, and producing materials in languages needed by their patron base. Some vendors take the library's already encoded EACC for Asian languages, as an example, and convert it on the fly to Unicode as users request particular records; others convert EACC wholesale to Unicode.

As with all of the Tech Notes, the links for Unicode in the text and in the bibliography are especially crucial, as they supply further technical information and direction.

Bibliography

The Unicode Standard, Version 3.0, Addison Wesley Longman, 2000. Hardcover, with CD-ROM. ISBN: 0-201-61633-5 ca 1070pp.

The Unicode Standard: A Technical Introduction
This is certainly technical, but it is understandable by anyone with a basic grasp of how computers work.

There are several online discussion groups for Unicode.

Hoffman, Paul, “Bringing in more of the world with Unicode” in *Network World*, March 27, 2000. p49.

Jimmy Thomas of CARL, Bob Rasmussen of Rasmussen Software, Inc., and Joan Aliprand of RLG contributed to this Tech Note with telephone and email advice. Any errors of course are mine own.

The Public Library Association's Tech Notes project grew out of the desire to continue the work of *Wired for the Future: Developing Your Library Technology Plan* by Diane Mayo and Sandra Nelson, published for PLA by ALA in 1999. Each of the Tech Notes, written by GraceAnne A. DeCandido, is a Web-published document of 1500-2000 words, providing an introduction and overview to a specific technology topic of interest to public libraries at a particular point in time. Topics were identified by PLA's Technology in Public Libraries Committee. Each Note is marked with the date of its completion and posting, and updates are noted.
The Technology in Public Libraries Committee is currently evaluating if the Committee should request PLA funding for additional Tech Notes. Readers' comments and suggestions are welcome and should be addressed to pla@ala.org. Please use *Tech Notes* in your subject line.

Prepared by GraceAnne A. DeCandido for the Public Library Association, June 2000. ladyhawk@well.com

Return to PLA home page

Return to Tech Notes Index page

E-Books: I Sing the Book Electric

Geographic Information Systems (GIS): Mapping the Territory

Video Teleconferencing: Here, There, and Everywhere

Metadata: Always More Than You Think

DOI: The Persistence of Memory

Electronic Statistics: Counting Crows

Wireless Networks: Unplugged, and Play

Intranets: The Web Inside

Push Technology: Pushed to the Brink

Digital Disaster Planning: When Bad Things Happen to Good Systems