Unicode

By Richard W. Boss

Unicode is an international character-encoding standard designed to support the electronic exchange, processing, storage, and display of the written texts of all of the world's languages and scripts, including scientific, mathematical, and technical symbols. Unicode is an encoding standard, not a software program or a font. The standard is independent of operating systems, programming languages, database management systems, and hardware platforms. Before Unicode, there were hundreds of different encoding schemes; no single one could contain enough characters. The European Union alone required several for its languages.

The Importance of Unicode for Libraries

While academic libraries have had holdings in a number of languages and scripts for many decades, public libraries—as the result of the increasingly diverse communities that they serve-- have begun to purchase books in a wide variety of languages and scripts. There is hardly a metropolitan area in North America that does not have immigrants from a score of countries. Even public libraries in small communities are serving an increasing number of people for whom English is not their primary language.

Unicode is also important for resource sharing. Libraries need to be able to borrow or lend materials in a wide variety of languages and scripts on behalf of patrons who cannot find what they need in their own libraries. The ability to send and receive bibliographic information among Unicode-conforming libraries is a major breakthrough.

Even libraries that have no interest in Unicode for themselves should appreciate the fact that vendors of integrated library systems and other information technologies that support Unicode are able to market their products globally, therefore, strengthening them financially and making it possible to maintain aggressive product development programs.

ASCII and Unicode

Before Unicode, the most widely used character set in the United States and a number of other countries was ASCII (American Standard Code for Information Interchange) or ISO/IEC 646, an international standard virtually identical to ASCII. ASCII's 8-bit

architecture [actually seven, with the eight a parity bit for error checking) supports just 128 characters (32 of them control characters), therefore, a binary number stored in a computer to represent a character in one language may be linked with a different character in another language. Its limited capacity means that only a fraction of the world's language and scripts can be supported.

Unicode, which became ISO/IEC Standard 10646 in 1991, provided 65,000 numbers to represent characters, enough to accommodate all of the world's languages and scripts except those that use ideographic characters. Version 3.1 subsequently expanded the capacity of Unicode to accommodate more than 70,000 ideographic characters. Version 5 is the current standard.

ASCII has not disappeared; however, it has emerged as UTF-8, (Unicode Transformation Format), the first 8-bit byte of Unicode, in Unicode Version 4 in 2003. That is all that is needed for English and several other languages. A plain ASCII string is also a valid UTF-8 string; no conversion needs to be done for existing ASCII text.

Two bytes are needed for Latin letters with diacritics and for characters from Arabic, Cyrillic, Greek, and Hebrew languages. Three bytes are needed for the almost all of the rest of the world's languages and scripts. Four bytes are available in Unicode, but rarely used in practice.

How does Unicode Work?

Unicode is designed to help programmers create software applications that work with any language and in any script Unicode avoids the time-consuming and costly task of separately developing the character set for each language and script, and maintaining similar, but separate source codes for each language and script. Instead, the programs are written in the vendor's language of choice and then internationalized using Unicode. It is still necessary to "localize" the programs for each language, but that is done without affecting the source code. An aspect of localization is supporting the right to left writing pattern of some languages. In that regard, it is interesting to note that a computer system client that is used to create records that include both English and Hebrew or Arabic, must provide a "virtual keyboard" as an alternative to the standard English keyboard and must automatically change the direction of writing when the vernacular is being entered.

Unicode support divides up into two categories: server and client. The server deals with storage; the client deals with displaying, printing, and editing. Full Unicode-compliance requires that both categories are supported.

The Unicode Consortium

The Unicode Consortium (www.unicode.org) is a non-profit organization that was founded to coordinate the development and promote the use of the Unicode standard. Its goal is to make Unicode the universal character encoding scheme, replacing schemes that cannot accommodate all of the world's languages and scripts. The membership of the organization consists of a broad spectrum of corporations and organizations in the

computer and information technology sectors. It cooperates closely with many other international standard setting organizations.

General Support

Microsoft has been a leader in the move to Unicode. Its Windows products are built on a base of Unicode, thus making it possible to market the products worldwide with minimum modification. Microsoft's first application of Unicode was with its introduction of Windows into East Asia. It was less expensive to use Unicode than to develop Chinese, Japanese, and Korean versions separately. Microsoft subsequently folded in Middle East and South Asian support. AIX, Solaris, HP/UX, Unix, Linux, and MacOS responded with Unicode support to remain competitive. Almost all DBMSs, including Oracle, Sybase, and DB2, now support Unicode. All the Web standards, including HTML and XML, support Unicode, as do Internet Explorer, Mozilla Firefox, and Netscape browsers.

OCLC and LC Support

OCLC introduced partial support for Unicode in the second quarter of 2002, but it was not possible to export vernacular records using Unicode until the fourth quarter of 2006. OCLC uses UTF-8 for English and several other languages because the first 128 bits in Unicode are ASCII. It is constantly expanding the use of Unicode to more languages.

The Library of Congress implement Unicode in its Voyager database in 2005 and included vernacular records in Unicode in its bibliographic records service for several languages and scripts not long thereafter.

Integrated Library System Support

VTLS' Virtua was the first integrated library system to fully comply with the Unicode Standard. In March of 1999, the vendor announced that all data in all records were being stored in the Unicode encoding scheme, thus allowing users to catalog and access records in their local languages. In addition to storing all data in the Unicode character set, Virtua was designed to support direct input, indexing, and display of characters in Unicode from a single, standard workstation. A user, whether a cataloger or a patron, can dynamically change the interface language and search language without affecting anyone else on the system. Another feature of the Virtua implementation of Unicode was a translator that converted all records not already encoded in Unicode to the Unicode character set. The company's commitment to Unicode has been a major factor in its global sales success. Nearly two-thirds of its installations are outside North America.

Ex Libris, which began with Hebrew as its software development language, but which had to look outside its home country for a larger market, was also an early adopter of Unicode. Innovative Interfaces' entry into the East Asian market led it to implement Unicode after initially seeking to develop the various language versions separately. Of these vendors, Innovative, which has a large number of public library customers, found that many of its public library customers had materials in their collections that required

Unicode to properly display the records for them.

By 2004, most major vendors of integrated library systems were supporting Unicode 4.0, the then current version of the standard. As of the first quarter of 2008, most of these vendors were supporting Unicode 5.0. However, there are still vendors, especially vendors of small systems and vendors that do not sell outside North America that do not support the Unicode 5.0 standard.

Specifying Unicode

Libraries that want to include records for many languages and scripts in their catalogs so that their users will have access to records in the vernacular, rather than in transliteration, should specify full Unicode 5.0 conformity on both the server and client sides. Libraries should require vendors that do not support Unicode 5.0 as of the date of their response to quote a future general release date for Unicode 5.0 conformity.

It is not necessary to specify UTF-8 as Unicode 5.0 incorporates it; nor is it necessary to specify ASCII support.

Sources of Information

The best source for both general and technical information about Unicode is the website of the Unicode Consortium at www.unicode.org

The full text of Unicode 5.0 is available as a monograph entitled The Unicode Standard, Version 5.0, Addison-Wesley Professional, 2006. (ISBN 978-0-321-48091-0).

Vendors of automated library systems are also good sources of information as even those who do not yet support Unicode have staff members who are knowledgeable about the standard.

Final version, May 8, 2008

Copyright Statement | Privacy Policy | Site Help | Site Index

ALAAmericanLibraryAssociation
50 E Huron St., Chicago IL 60611 | 1.800.545.2433