Unicode

By Richard W. Boss

Unicode is the key to building multi-lingual databases. The protocol, which is international, is generally defined as a unique number for every character no matter what the hardware platform, software program, or language.

Now that a many of the major vendors of automated library systems are selling systems throughout the world, it is important to them to support far more characters than the ubiquitous ASCII character set can accommodate. Academic research libraries that collect globally also must be able to include records for many different languages and scripts in their catalogs. The increasing diversity of North America's population also makes it increasingly important for its public libraries to acquire, catalog, and make available materials in many languages, and to store and display the records in the vernacular, rather than in transliterated form.

ASCII 's 8-bit architecture supports just 256 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. Unicode, which became ISO/IEC Standard 10646 in 1991, provides 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. The current version of the standard is 4.0.

How does Unicode Work?

Unicode is a standard for encoding characters, rather than a software program or a font.

It is designed to help programmers create software applications that work with any script and in any language in the world. It is independent of the operating system, database management system, or hardware platform employed, however, each of these must accommodate the standard. The standard is promoted by the Unicode Consortium, a non-profit organization open to any organization or individual who is willing to support the standard.

Unicode avoids the time-consuming and costly task of separately developing the character set for each script and language and maintaining similar, but separate source codes for each script and language. 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 cataloging 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 have been completed.

General Support for Unicode

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 merged

in Middle East and South Asian support. AIX, Solaris, HP/UX, and MacOS responded with Unicode support to remain competitive. Almost all DBMSs, including Oracle, Sybase, and DB2, support Unicode. All the Web standards, including HTML and XML, support Unicode, as do Netscape Navigator and Internet Explorer.

Automated Library System Support for Unicode

VTLS' Virtua was the first automated 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 converts 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 60 percent 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. Endeavor Information Systems, a vendor that counts many academic research libraries among its customers, has also implemented Unicode. Innovative Interfaces' entry into the East Asian market led it to implement Unicode after initially seeking to develop the various language versions separately. Dynix's Horizon, which is used by a number of academic research libraries, has also adopted Unicode.

Sirsi was still working on Unicode as of mid-2003. It had completed the client side, but expected to take up to 12 months more to complete the server side. BiblioMondo's Portfolio, GIS' Polaris, and Geac's VUBIS were only partially completed as of mid-2003. TLC has been marketing a Unicode-compliant cataloging client called ITS.International for more than two years, but it had not fully implemented Unicode support as of mid-2003.

Specifying Unicode

Libraries that want to build and maintain multilingual databases with records for other languages and scripts in the vernacular should specify Unicode 4.0-conformity on both the server and client sides. If Unicode 3.2 is the latest version supported, the vendor should quote the scheduled date for implementation of Unicode 4.0. Libraries should be prepared for vendors that did not support Unicode as of mid-2003 to quote general release as long as 12 months from the time the response to the RFP is submitted.

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 4.0 is available as a monograph entitled The Unicode Standard, Version 4.0, Reading, MA: Addison-Wesley, 2003. (ISBN 0-321-18578-1).

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

The Technology for 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.

PLA Tech Note by GraceAnne A. DeCandido