The Story Thus Far

In my previous column I traced the development of character sets in computers from the fundamental coding of ones and zeroes to the founding of the first universal character set, Unicode. We are currently in a transition from the functional but awkward sets of code pages that allow computers to form characters in different scripts to a single, universal character set to represent all languages on computers. Like the progression from block printing to moveable type, these advances in computer technology have reduced the effort and cost of producing written materials throughout the literate world. The question for us today is: How does this affect libraries?

Libraries and Languages

Although some individual library collections may be limited to works in a single language, the overall library space is definitely multi-lingual. There are some special difficulties in fulfilling the language needs of library catalogs. To begin with, libraries need to handle the full range of characters in a single system. This is different to the need of writer or a publisher who produce documents generally in one language, or possibly two languages, per document. For those publications, the use of special type for printing or of code pages for digital applications would usually suffice. Furthermore, library catalogs are not just printed expressions: they must be searchable by library users; they must be suitable for sorting in a variety of orders; they may be downloaded into bibliographic software packages; and they may be exchanged and edited by library staff. The active nature of library data means that the system must manage all of these functions across all of the languages in the library.

The MARC record provides for a range of accented characters through its own redefinition of the 256 bytes available for character expression. This definition preceded the computer world's use of code pages and responded to a different problem set, in particular the need to intermingle different scripts in a single character set so that a viewable page of retrievals could display records representing works in a variety of languages. MARC manages to incorporate a large number of characters within the 256 bytes by defining letters and diacritics separately rather than as combined characters. For example, MARC has the vowels a, e, i, o, and u, and it has accent marks such as the umlaut, grave, and acute. These eight can be combined to make twenty four different accented characters, from â to ú. MARC uses some remaining bytes for characters that can't be expressed as a character/diacritic pair, such as the Icelandic thorn or the slash o of the Scandinavian languages. MARC also defines a few characters for its own special needs, such as the subfield code character, which is often expressed as a dollar sign, as in "$a."
As the capabilities of both computer displays and of printing from computers progressed, new methods were added to the MARC record and the cataloging systems used by libraries that allow libraries to encode some non-Latin based alphabetic languages, such as Greek or Russian, as well the ideograms of Chinese, Japanese, and Korean. In libraries in the United States, these non-Latin characters are generally encoded in special fields in the MARC record (the 880 fields called "Alternate Graphic Representation"), while Latin-based transliterations still occupy the primary descriptive fields for authors and titles. This permits display of the vernacular language of the works while the search and sort functions make use of the transliterated form of the headings.

This describes the library catalog as most of us know it today. Currently available library system software allows libraries to input and display Western European languages and the following non-Latin scripts: Arabic, Chinese, Cyrillic, Greek, Hebrew, Japanese, and Korean. Now with the advent of Unicode, it will be possible for library catalogs to display all works in their vernacular script. However, it also brings up a number of questions that need to be answered before this conversion takes place.

**Catalogs and Unicode**

Unicode promises to transform library catalogs into truly multi-lingual databases, as rich as the books on the library shelves. Before we can embrace this new character set, however, we have to understand how it will interact with catalog creation and use.

**DISCOVERY:** How does Unicode interact with the abilities of library users and the tools that we provide for them: databases, keyboards, screens? What does it mean for the ordering of records in displays?

**CATALOGING:** What effect will the expanded character set have on cataloging and on the training of cataloging staff? Does metadata have language requirements that differ from the literary uses of written language?

**TRANSITION:** What is the best way to transition from the MARC-8 character set in use today to a Unicode-compliant system?

**Discovery in the Unicode Catalog**

The purpose of public catalog records is to facilitate the discovery of works held in the library by library users. In the card catalog, discovery took place through the alphabetical listing of headings, primarily for authors (and other creators), titles, and subjects. In the online catalog, discovery requires a user to type on a keyboard either the beginning of a heading, or one or more keywords to be found in the catalog’s bibliographic records. Items retrieved through an online catalog search are generally presented either in alphabetical order (by main entry) or date order. With the prospect of making use of Unicode, both the typing of queries and the display of results become problematic.

If you are a library with works in Greek, and you can assume that users searching for these books are conversant with the Greek language, you can provide workstations with either analog or virtual Greek language keyboards, so that your users can perform a search for the author ΑΛΚΥΟΝΗ ΠΑΠΑΔΑΚΗ. But suppose that yours is a library serving mainly an English-speaking population, and a user in your library has a citation to this book:
A computational interpretation of the \(\lambda\mu\)-calculus / by G.M. Bierman

This user will need to be using a standard U.S. keyboard to type most queries, and yet in this case has a need for Greek characters. It is also likely that the user will not be conversant with the Greek language or keyboard, only with a few Greek characters used as symbols in mathematics. If the remainder of the title is distinctive, the user can search only on the words using the Latin alphabet and may still be able to retrieve the item. This does, however, beg the question of the sort order of the display. For the book above the Greek letters appear near the end of the title heading, where would you find (or expect to find) the book titled \(\lambda\)-calculus and computer science theory in a browse display? And where would works by ΑΛΚΥΟΝΗ ΠΑΠΑΔΑΚΗ appear in relation to the Latinized version of her name, Alkioni Papadaki? There is no sort order that would intermingle the different Unicode scripts, so in Unicode-based catalogs today, browse displays can progress from Greek to Russian to Hebrew:²

- \(\Sigma\)μβολή στη βελτιστοποιηση της εκχυλισης του διασπορικου βωζτ
- Έκαψη Μαρσ = Jakšic : March ; Ποζδραβλαν τε. : Μαζυρκα.
- Καρανφιφικε Karanfilje ; Српски потпури = Srpski potpuri
- српски потпури.
- Έκαψη Αντι Αλαρ : Χολεστερόμο, Υγεία Υπολογιστής : Υονίμο Βσηρά
- Απολογείσα Μεντι Τέρι, Ρουμέι Ρούβρυς : Σκοπός Αριτμ.

Would our user expect to find \(\lambda\)-calculus and computer science theory in among the works in Greek? Probably not.³

Not only are the non-Latin characters a challenge for sorting, so are some of the Latin characters. For example, in Swedish, the ö sorts after the z, while in German it sorts after the o. For this reason the Unicode standard does not provide a single sort order, but advises: "It is important to ensure that collation meets user expectations as fully as possible."⁴ This means that a Swedish library would sort the ö after the z, a German library would sort it as oe and a U.S. library may decide to sort it simply as another letter o. For the non-alphabetic⁵ languages like Chinese, order across the many thousands of characters, complex rules determine order. Even more problematically, some Unicode characters, such as the symbols ◊ or ↑ have no known order. Like the famed "artist formerly known as Prince" we will need some agreement on how we will treat these types of symbols for searching and sorting.

The technical decisions regarding the implementation of Unicode in library catalogs would be much easier to make if we could guarantee that all of our data would be consistent in its use of characters. Catalog records created prior to Unicode will have taken the option of spelling out any characters or symbols that the cataloger could not render in the current character set. A title with \(\lambda\)-calculus will have been entered as \([\text{lamba}]-\text{calculus}\). This title will sort under the Latin letter "l." Any decisions made about the use of Unicode will have to take into account that library catalogs contain legacy data from the pre-Unicode era.
Cataloging in Unicode

The only mention of scripts in the cataloging rules instructs catalogers to use what scripts are available when transcribing descriptive data. There is also the "language of the catalog" which determines the language of subject headings and which may also determine whether the library chooses to transliterate headings in non-Latin scripts. To date, the scripts available to catalogers have been those included in the MARC21 character set, called "MARC-8." This character set is not only much smaller than Unicode in its number of characters, but rules for its use has been incorporated into the input functions of library utilities, such as OCLC and RLIN, and library vendor systems. A question that the library community is struggling with today is whether the library use of Unicode should have similar library-specific rules, or whether libraries should make use of the whole of Unicode, even where that character set offers more than one option for representation of a character.

Although the early developers referred to Unicode as "universal and uniform", the reality of written language means that a single solution to script representation was not always possible. An example that we are familiar with in the library world is that of characters with accent marks. In countries where the language contains accented letters, writing tools like typewriters and computer keyboards contain keys for the accented letters, not separate keys for letters and accents that need to be combined during writing. These language groups developed computer character sets, such as the ISO 8859 standards, that contained the accented letters as characters in themselves. Other communities, such as the U.S. library community, use separate letters and diacritics. Unicode does not make a choice between these two methods, but provides characters for both options.

Furthermore, because Unicode defines characters based on their meaning, not their display, some display characters appear more than once in Unicode, but with different meanings. As an example, the Greek Ω (omega) is both a letter in the Greek alphabet and a symbol in the set of special characters for mathematical notation, and the Å is both the upper case Latin A with a circle and the scientific notation for angstrom. In addition, similar punctuation marks appear in various parts of the Unicode standard, with differences in meaning that are significant to language experts but may be elusive to most viewers of the text. These subtleties will be a challenge for the catalog's goal as a search technology for library users, as well as to the training of catalogers.

The MARC21 Transition to Unicode

In 1998, a modification was made to the MARC21 format to allow records to be transmitted in Unicode. Because the MARC21 record standard is a communications format (that is, it defines how records will be communicated from one system to another) there was an important question of compatibility between Unicode and MARC-8. If some libraries were to embrace the entirety of Unicode, how could they exchange MARC21 records with libraries using the more limited MARC-8 character set? What would happen when a MARC-8 library received records with Unicode characters that could not be translated to MARC-8? As an interim measure, the MARC21 community decided that, for the present, Unicode would be allowed as a MARC21 character set but it would be limited to those same characters that can be expressed in MARC-8. This has not prevented some vendors who market their systems outside of the U.S. and Western European market from developing catalog software for users of languages outside of the MARC-8 group, but it has stalled the expansion of the MARC21 standard to a wider use
of vernacular expression of non-Latin languages. There are also differences of opinion among library systems vendors to whether it would be best to make a giant leap into full Unicode, or introduce languages and scripts gradually over time. For some, the change to Unicode would require expensive development and may not be of great benefit to their customers.

Unlike the difficult issues brought up regarding public access in a Unicode-based system, the technical change of catalog records from MARC-8 to Unicode is actually not a complex nor lengthy process. Translation tables that can be manipulated by computer programs are available on the Library of Congress web site, and the modern computer languages, such as C, Java, and Perl, are able to work directly with the Unicode character set, as are today’s web browsers. When moving to a new system, the translation from MARC-8 to Unicode becomes a step in the inevitable conversion of the old database to the new. In daily operations, incoming MARC21 records in MARC-8 format will be converted during the input process. These conversion steps are minor compared to other input processes, such as creating indexes, and may not visibly impact the library system in terms of processing time.

Conclusion

Although the first version of Unicode was issued in 1991, we are still not living in a Universal Character Set world. If you do a search for the Unicode term H₂O in Google Scholar, your retrievals consist of documents that contain the letter H near the letter O. A search on H2O, however, retrieves documents with the chemical formula for water expressed in its simple ASCII representation, H2O. In article databases covering the sciences, characters like λ are spelled out (“[lambda]”). This legacy will remain with us long after Unicode becomes the norm. Yet elsewhere on the Internet you can find entire web sites using Unicode as their character set.

Libraries are beginning the transition to Unicode as library system vendors offer the MARC21 Unicode scripts as the character set of the database and displays. Web browsers can detect that web pages are in Unicode and can display a wide range of scripts if the user’s computer operating system is Unicode-aware. If libraries make the leap to full Unicode use, as it appears they might, they will find themselves on the leading edge of this transformation. There will be many challenges, including working with legacy data alongside records in full Unicode. Will transliteration still be needed? What training will be required for library staff and library users? Unicode has the promise to provide a richer experience for information seekers, once we work out the details.
Code pages are character sets used by computer applications, with different code pages for different languages or language groups. See the previous article in this series for more information.

These examples were taken from the browse display of the MELVYL catalog, which uses Ex Libris' Aleph software. http://melvyl.cdlib.org

This illustrates how difficult it is to define the language of an individual heading, and that the language of a heading, such as for an English language work of literary criticism entitled "Les Misérables," can be different from the language of the work.


The MARC-8 character set is defined on the Library of Congress web site at:


The same search on Yahoo retrieves items containing the keyword "H2O." Yahoo appears to normalize the query to "h2o" while Google normalizes it as "h + o."

Unicode is a computing industry standard for the consistent encoding, representation, and handling of text expressed in most of the world's writing systems. Developed in conjunction with the Universal Coded Character Set (UCS) standard and published as The Unicode Standard, the latest version of Unicode contains a repertoire of more than 128,000 characters covering 135 modern and historic scripts, as well as multiple symbol sets. The standard consists of a set of code charts for visual reference, an The Unicode Standard is a character encoding system that defines every character in most of the spoken languages in the world. To overcome the limitations of existing character encodings, several organizations began working on the creation of a global character set in the late 1980s. The need for this became even greater with the development of the World Wide Web in the mid-1990s. The Internet has changed how companies do business, with an emphasis on the global market that has made a universal character set a major requirement. A global character set needs to fulfill the following conditions: Request PDF | On Jan 31, 2006, Karen Coyle and others published Unicode: The Universal Character Set Part 2: Unicode in Library Systems | Find, read and cite all the research you need on ResearchGate. We evaluate the Guichardet-Wigner pseudocharacter on the universal covering group of SU(1,1) explicitly and find the level sets of this pseudo character. Read more. Article. Unicode: The universal character set. November 2005 Â· The Journal of Academic Librarianship. K Coyle. Read more. Article. The Descriptive Character of Certain Universal Sets. October 1973 Â· Proceedings of the London Mathematical Society. D. G. Larman.