The Unicode Standard

Its Scope, Design Principles, and Prospects for International Cataloging

Joan M. Aliprand

The Unicode Standard is a global character set for worldwide computing covering the major modern scripts of the world as well as classical forms of Greek, Sanskrit, and Pali. The history and implications of Unicode Standard are discussed. The principles underpinning the design of the Unicode Standard are described with reference to those principles that also are present in USMARC and UNIMARC. Unicode give the potential to support every script. Expanding the character set would have consequences for transcription. Faithfulness of transcription has implications for retrieval. The addition of more characters to support more exact cataloging affects the economic cost of cataloging. The need for characters should be related not to the production of a surrogate for the physical item that has been cataloged, but to facilitating retrieval.

A common question about the Unicode Standard is "Is my script there?" The Unicode Standard covers the major modern scripts of the world and also classical forms of Greek, Sanskrit, and Pali. It includes more than 21,000 East Asian ideographs—7,000 more than the East Asian Character Code (EACC) used in USMARC (American National Standards Institute 1990)—and the full complement of modern Korean *hangul* (EACC has only a fifth of these.)

Figure 1 shows the content of version 2.1 of the Unicode Standard. The content of the entire code space is on the left. The enlargement on the right shows alphabetic scripts in more detail.

Version 2.1 of the Unicode Standard is made up of the published book, *The Unicode Standard*, version 2.0, augmented by Unicode Technical Report number 8 (published on the Web). The characters added in version 2.1 were the Euro currency sign and the object replacement character that marks the position of data that cannot be used in a plain text context.

The latest version, 3.0, includes 11 more scripts (Burmese, Canadian Syllabics, Cherokee, Ethiopic, Khmer, Ogham, Runic, Sinhala, Syraic, Thaana, and Yi), as well as symbols for Braille and an additional 6,000 East Asian ideographs.

Where Did Unicode Come From?

In 1988 Joe Becker of Xerox and Lee Collins and Mark Davis of Apple started to think about a better way to perform multilingual computing: a character set as simple and basic as ASCII that met the needs of the whole world. Becker called it "Unicode." Other companies joined the project. The Research Libraries Group (RLG) came in very early, because it developed the EACC.

Joan M. Aliprand (Joan_Aliprand@notes.rlg.org) is Senior Analyst at the Research Libraries Group (RLG).

The author wishes to thank librarians at Stanford University—Edward A. Jajko, Barry Hinman, Heidi Lerner, and John E. Mustain—for providing examples to illustrate the various cases.

©1998 by the Research Libraries Group, Inc. Unicode is a registered trademark of Unicode, Inc.

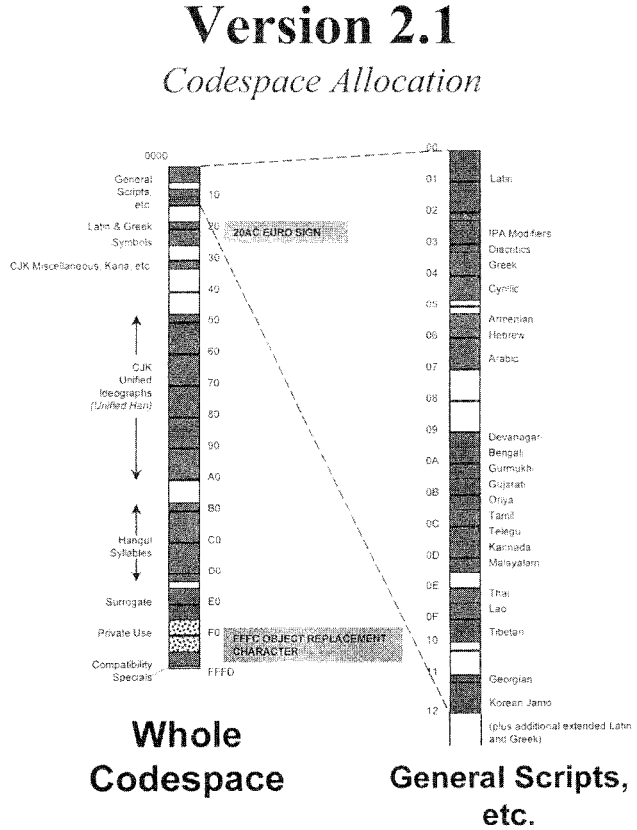

Figure 1. The Unicode Standard, Version 2.1

The Unicode Consortium, founded in 1991, is an international organization responsible for the development and promotion of the Unicode Standard. A list of full and associate members can be found on the Web at www.unicode.org/unicode/consortium/memblogo.html. Full members have voting privileges and so determine the content of the Unicode Standard.

Around the same time that work on the Unicode Standard began, Joint Technical Committee 1 (JTC 1) of the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) were also developing a global character set. The project's identification number was "10646."

JTC 1 has established procedures for the development of the standards for which it has responsibility. The content of ISO/IEC 10646 is determined by the representatives of national standards bodies, which have declared their intent to participate in the work (ISO/IEC 1993). The library and publishing part of ISO, Technical Committee 46 (TC 46), has no direct role, but can provide comments as a liaison organization.

In 1992 version 1.0 of the Unicode Standard and the second DIS (draft international standard) of ISO 10646 merged. Ever since then, the character repertoire and encoding of the Unicode Standard and ISO 10646 have been kept in sync. The difference between the two standards is that Unicode also specifies character properties and implementation rules that are required if applications are to be mutually consistent.

Who Is Using Unicode?

You'll still hear comments such as, "Unicode is all very well, but who is actually using it?" I think the question should be: "Who's not using it?" Let me drop a few names of Unicode users: Java from Sun, Windows NT and Internet Explorer from Microsoft, Netscape Navigator, database products from Oracle and Sybase, Mac OS 8.5 from Apple . . . the list goes on and on and grows constantly.

More significantly, standards are beginning to reference Unicode and ISO 10646. Have you looked at the specification for HTML 4.0 (W3C 1998a)? How about XML 1.0 (W3C 1998b)? The UTF-8 form of Unicode (Unicode Consortium 1996) has been endorsed as "best current practice" by the Internet Architecture Board (IETF 1998).

What Are the International Implementations?

Unicode can be used for any application—single script, multiscript, or fully global—so I'm not sure why international implementations should be singled out. The Web knows no boundaries either.

Library Issues

Now it's all very well having the prospect of Unicode and ISO 10646, but the key question is: How are we going to take advantage of it? And this leads to thinking about how we are going to use Unicode data in the machine-readable records that we exchange.

I stress "exchange" here. What you do internally is your own business. Indeed, there are library systems in operation today that use Unicode internally but do not send it out because the exchange formats do not yet specify how to do that.

Part of the task we are faced with is "how to get from here to there." The ultimate goal is to be able to use Unicode data in library records. But we have an enormous legacy of records encoded in 7- and 8-bit character sets that cannot be abandoned. One of the essentials is to define mappings between today's character sets and the character repertoire of Unicode/ISO 10646.

How Will Unicode Be Used in Library Records?

There are two families of MARC records today: USMARC and its derivatives, and UNIMARC and its offspring. Each MARC defines the character sets that are legitimate for its records (Network Development 1994; Holt and McCallum 1994).

ALA's MARBI Committee, which advises the Library of Congress (LC) on the USMARC formats, delegated work on the use of Unicode in USMARC records to its Subcommittee on Character Sets and special task forces.

Mappings have been defined for all the single-byte character sets and approved by MARBI. They were published on LC's Web site (*USMARC to Universal Character Set Mappings* 1998).

Mapping of the multibyte EACC is underway. Ideographs are being finalized. Drafts for Japanese *kana* and Korean *hangul* have been prepared. When the work is complete, a proposal will be submitted to MARBI.

Proposal 98-18: *Unicode Identification and Encoding in USMARC records*, prepared by the MARBI Unicode Encoding and Recognition Technical Issues Task Force (1998), was on the agenda for MARBI's June 1998 meeting. A key recommendation in this proposal was the use of UTF-8.

Other issues are still open, and will be addressed by MARBI. For example, Discussion Paper 111 (1998) considered continuing use of the 880 field, *Alternate Graphic Representation*.

A similar mapping process for the character sets used by European libraries (including those specified for UNIMARC) was undertaken by the CHASE committee, part of the European CoBRA project (Fisk and Brickell 1997; CoBRA+ Computerised Bibliographic Record Actions 1998).

In both the MARBI project for USMARC and the CHASE project, compromises had to be made in defining the mappings. Perfect round-trip mapping for every character is not possible unless values in the Private Use Area are utilized.

What about Z39.50?

And what about Z39.50? Version 3 of Z39.50 provides for character set negotiation between the origin and target systems (*Character Set and Language Negotiation* 1998). The system-to-system negotiation is based on ISO/IEC 10646 conventions, but does not appear to have incorporated its more recent amendments. The specification for character set negotiation does apply only to the protocol, and does not determine whether a record is transmitted (on the basis of its character content). The issue of unknown or undisplayable characters is discussed below.

How Should I Sort Multiscript Data?

Another topic of great concern to librarians is sorting. The ISO working group that deals with internationalization of computer systems has been developing a default ordering for the character repertoire of ISO 10646; the latest version is designated "ISO FCD (Final Committee Draft) 14651" (ISO/IEC JTC1/SC22/WG20 1997). Unfortunately, there are defects in FCD 14651; for example, rules for some scripts are lacking.

The Unicode Consortium has published a draft of the Unicode Collation Algorithm to provide a complete, unambiguous, specified ordering for all characters in the Unicode Standard, version 2.1 (Davis and Whistler 1999). The consortium is actively involved with the ISO effort through its membership in ANSI/NCITS Technical Committee L2.

An aspect of "how to get from here to there" is defining Unicode equivalents for the characters we currently use. The two main sources of character sets specifically for library use are LC and TC 46. LC specifies character sets to be used in USMARC records. TC 46 has developed various character sets for bibliographic data exchange. UNIMARC specifies use of certain ISO character sets developed by TC 46.

This does not mean that a particular MARC format can only use USMARC or TC 46 character sets. The character sets to be used in a particular MARC format are determined by the specification for that format, and can dictate use of another character set (e.g., a character set widely used in a particular country). USMARC and TC 46 Unicode and existing library character sets are discussed here because they are of international interest.

The Unicode Standard and ISO 10646 do not include every character encoded in library character sets. This has become apparent in the MARBI and CHASE work on use of Unicode in library records.

The MARBI work on mapping between the current USMARC character sets and Unicode identified seven additional Arabic script characters. These have been accepted by the Unicode Technical Committee and are proceeding through ISO balloting. At the time of writing, their code assignments were tentative. They were outside the scope of version 2.1 of the Unicode Standard, but are in version 3.0.

Differences between EACC and the Unified ideographs of Unicode/ISO 10646 require Private Use values if the integrity of certain EACC characters is to be preserved, although EACC contains a unified set of ideographs that can be used for Chinese, Japanese, and Korean (CJK).

Working Group 1 (WG1) of Subcommittee 4 of TC 46 is responsible for library character sets. The TC 46 character sets specified for use in UNIMARC are listed in table 1. WG1 has identified a number of characters in its character sets that are not in the Unicode Standard or ISO 10646. The

Irish National Standards body has proposed their addition to ISO 10646. The Unicode Technical Committee and Technical Committee L2, the United States body that deals with coded character sets, considered them when they met in July 1998. (RLG is a member of L2, and a voting member of the Unicode Technical Committee.)

The Unicode Standard and ISO 10646 both include a private use area that has to be used when the integrity of specific characters cannot be compromised. The library community needs to coordinate its use of private use values: this has been suggested for European libraries through the CHASE project. But library data is exchanged globally, so it needs to be done not just independently for a particular group of users, but on a worldwide basis.

User Concern: Doesn't 16 Bits Mean Doubling of Disk Space?

The next set of questions addresses concerns that surface quite regularly. One worry that comes up again and again has to do with disk space: 16 bits is twice the size of 8 bits, so won't disk space requirements be doubled?

This question ignores the technical side of things. If UTF-8 is used as the encoding form, the space requirements will be different and will relate to the character content of your data. If the Unicode Compression Algorithm is applied, space requirements will be lessened (Wolf et al. 1998).

But what I find strange about this question is that no one seems concerned about the far larger quantity of disk space needed for digitized images and multimedia. And the hard bottom line is: like it or not, library vendors are moving to Unicode. You either stay with what you have forever, without the scripts you want; or, at some stage, you upgrade to a new system that does the scripts you want and you pay the going price.

User Concern: How Will 16 Bits Affect the Speed of Data Transfer?

Another worry is that 16 bits will affect the speed of data transfer. If you've downloaded any multimedia clips, you'll find that images, sound, and video are the bandwidth hogs, not textual data. Furthermore, UTF-8, where all Basic Latin characters are represented in 8-bits, is recommended for the Internet.

Table 1. TC 46 Character Sets Specified for Use in UNIMARC

Identification	Name
ISO 646	Basic control set and basic Latin set
ISO 6630	Bibliographic control set
ISO 5426	Extended Latin set
ISO Registration #37 (revised 1983)	Basic Cyrillic set
ISO 5427	Extended Cyrillic set
ISO 5428	Greek set
ISO 6438	African language set

User Concern: How Do I Cope with Unknown Characters?

When we talk about the problem of unknown characters, we need to take the scope of the problem into account. The inability to see a run of text (perhaps the bulk of a record) is a lot worse than the occasional unknown character (possibly caused by an encoding error). USMARC addresses the inability to see a run of text by appending the original script (in 880 fields) to a completely romanized record, so that a system lacking multiscript capability can at least display the romanized equivalent. Any system should have a graceful way to cope with an undisplayable character; each system will have its own convention.

Inability to display a character has two possible causes: the system has no information or it lacks the tools to display a known character

If the system has no information, either the character code value has not yet been assigned to represent a character, or it is a value in the private use area and the system has no information about the private agreement establishing a meaning for that value.

The other cause is not lack of information about the character but lack of the tools to display it. Most commonly, this is due to lack of a font. But just having a font does not guarantee satisfactory presentation for some scripts. Complex scripts such as Arabic or Devanagari require rendering software in addition to a font with an adequate collection of images (which is much larger than the number of characters for the script shown in the Unicode Standard).

How do we want to cope with such a situation? Should there be a requirement to code script information in the record so that a user can be forewarned about the need for script support? Or should an error message be presented instead of the record? Should undisplayable characters simply be presented as mysterious boxes? Should their Unicode value be shown to identify them?

The Unicode Consortium has designed a font containing a typical character for each script, to give a little more information about an undisplayable character. There might not be an elegant solution to this situation, but we need to consider the tradeoffs between the options, and, in particular, whether a strategy imposes any additional work on the cataloger.

Design Principles of the Unicode Standard

To provide a better understanding of the Unicode Standard, I'd like to illustrate some of the principles underpinning its

design (shown in figure 2). If you understand the design principles, you can understand why some things that you would call "characters" have not been coded.

Quite a few of the Unicode design principles are also present in USMARC and UNIMARC.

Characters, not glyphs. The USMARC Arabic set encodes the conceptual letters, not the positional forms used to write Arabic (Network Development 1994).

Plain text. Both MARC formats are for minimally legible data. Neither allows for "rich text" features such as typeface, font size, color, and so forth.

Logical order is used in USMARC, which is important when scripts run in opposite directions, for example, English mixed with Arabic.

Unification is exemplified in the EACC, developed by RLG and later adopted as a U.S. standard. In fact, unification applies to character sets for most languages: Basic and Extended Latin, Basic and Extended Cyrillic, USMARC Hebrew, USMARC Arabic. (In USMARC, Basic Latin is ASCII [ANSI X3.4] and Extended Latin is ANSEL [ANSI/NISO Z39.47]. In UNIMARC, Basic Latin is ISO 646, the International Reference Version of ASCII, and Extended Latin is ISO 5426. Both USMARC and UNIMARC specify the same Cyrillic characters sets: Basic Cyrillic is ECMA Registration No. 37, and Extended Cyrillic is ISO 5427).

LC developed *dynamic composition* of accented letters in the 1960s (Rather 1968).

Of the other design principles, the most interesting is *semantics*. Because Unicode characters have well-defined semantics, they have defined properties. For example, the properties of character U+0663, ARABIC-INDIC DIGIT THREE include its decimal digit value 3 and the bidirectional category *arabic number*.

The principle of semantics and the definable properties that devolve from this principle are what distinguishes the Unicode Standard from its counterpart, ISO 10646, Universal Character Set. Both publications are consistent in character repertoire and code point assignments.

ISO 10646 is a normal international character set. It defines characters and the codes which represent them, but says nothing more. Implementers need additional information about the character properties to produce consistent software. The Unicode Standard includes this information.

Software that conforms to the Unicode Standard is required to support normative properties. So with respect to USMARC and UNIMARC, it should be assumed that platform software used for library applications will include rules based on Unicode properties.

Another Unicode principle that needs examination in the context of international cataloging is *characters, not glyphs*. One's first reaction is to think about characters as equivalent to the exact glyphic form, because that is what we see and it also agrees with the tradition of the catalog entry as a surrogate for the item being cataloged. But that isn't always the case.

Sixteen-bit character codes
=> Unicode character codes are 16 bits. (Not two bytes, but an indivisible 16 bits.)

Full encoding
=> The entire codespace is available to encode characters.

Characters, not glyphs
=> The Unicode Standard encodes conceptual characters, rather than the elements of text that we see.

Semantics
=> Unicode characters have semantics, that is, name, representative glyph and normative properties.

Plain text
=> Plain text is a pure sequence of character codes.

Logical order
=> Logical order underpins the correct presentation of text. (For most alphabetic scripts, it is equivalent to the keystroke order of a perfect typist.).

Unification
=> Unification of characters across languages avoids duplicate encodings.

Dynamic composition
=> Creation of accented forms from a sequence of characters. (Libraries have done this since the beginning of automation.)

Equivalent sequence
=> A precomposed form maps to a sequence of other characters.

Convertibility
=> Character identity is preserved for interchange with a number of widely-used standards.

Figure 2: Design Principles of the Unicode Standard

A couple of examples will help explain the distinction between *character* and *glyph*. The lower case form of the Latin letter "g" can be written in one of two ways: with one bowl or two (as shown in figure 3). We are perfectly satisfied using the two forms interchangeably, depending on the type design we happen to be using.

Arabic letters can have up to four positional forms (as shown in figure 4). The letter's shape depends on where it is relative to spaces and other letters. Unicode encodes the underlying conceptual letter; the correct glyph for display can be determined by an algorithm.

The important point to remember is that what is encoded in the data doesn't necessarily correspond exactly to what you see as text. And this applies to USMARC and UNIMARC as much as to Unicode.

Unicode gives us a great many more characters than we have had hitherto in any cataloging environment. While not all platforms support every script properly yet, use of Unicode gives the potential for this support. I want you to think about how this new development interacts with transcription, an essential part of cataloging, and what the consequences would be if we expanded the character repertoire to provide for ever more exact transcription.

Back in the days of the unit catalog card (which some of us still remember) it was possible to write in a letter or symbol that was not available on our typewriters. The underlying principle was faithful transcription, that is, the catalog entry that stood for the work should reproduce the information from the work as exactly as possible.

We compromised on this when we automated. Until now, most online catalogs have been limited to Latin script. Now our immediate reaction is to ask for the addition of all the various typographical things that we see to the universal character set.

But before we do this we should stop and think about what is truly necessary, and if what we are proposing conflicts with Unicode design principles. We need to bear in mind that we have always made exceptions to exactitude, even when the typographical facilities were available.

We have to compromise when the text that appears on a title page is extremely long, as in the eighteenth century work shown in figure 5.

We also have to compromise when transcription is impossible, as in the mathematical formula in figure 6.

We don't transcribe all the ligatures and other calligraphic flourishes that are found in scripts such as Arabic, but replace them with regular letter-forms. The book shown in figure 7 is in Turkish, which was written in Arabic script in the days of the Ottoman Empire.

LC practice is to always transcribe Hebrew unvocalized, even when vowels and marks of pronunciation (which are positioned on consonontal letters) appear on the source of information. Figure 8 shows a Hebrew translation of Longfellow's poem *Evangeline.*Vocalization can be seen on the author's name (above the line) and the title (in the middle). The information at the foot of the title page is unvocalized. So we've never been 100% faithful.

Faithfulness of transcription has implications for retrieval. This was true in the days of the card catalog. What was the alphabetical order for that strange letter or symbol we inked in on the unit card? And how did a user know where we had filed it?

Ideographs provide an excellent example of the conflict between assigning a unique encoding to every glyphic form and how that affects retrieval. The character for "longevity" can be written in more than one way. Indeed the hundred forms of *shòu* that were used in antique writing is still a motif in Chinese art (see figure 9).

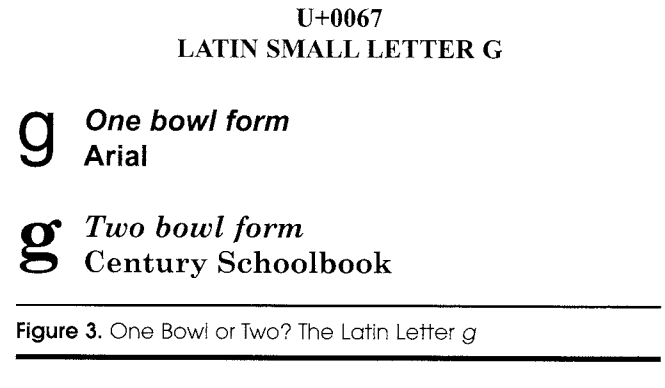

Figure 3. One Bowl or Two? The Latin Letter *g*

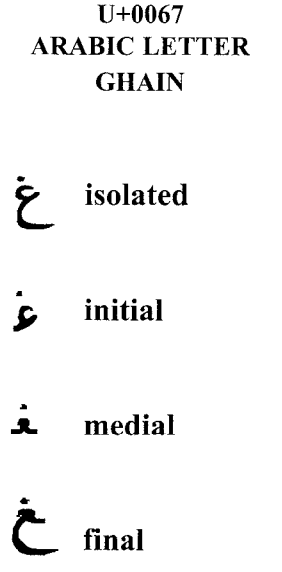

Figure 4. The Positional Forms of the Arabic Letter *ghain*

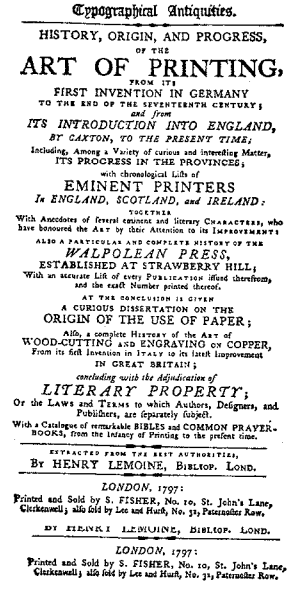
Typographical Antiquities.
HISTORY, ORIGIN, AND PROGRESS, OF THE
ART OF PRINTING,
FROM ITS
FIRST INVENTION IN GERMANY
TO THE END OF THE SEVENTEENTH CENTURY;
and from
ITS INTRODUCTION INTO ENGLAND,
BY CAXTON, TO THE PRESENT TIME;
Including, Among a Variety of curious and interesting Matter,
ITS PROGRESS IN THE PROVINCES;
with chronological Lists of
EMINENT PRINTERS
In ENGLAND, SCOTLAND, and IRELAND:
TOGETHER
With Anecdotes of several eminent and literary CHARACTERS, who have honoured the ART by their Attention to its IMPROVEMENT:
ALSO A PARTICULAR AND COMPLETE HISTORY OF THE
WALPOLEAN PRESS,
ESTABLISHED AT STRAWBERRY HILL;
With an accurate List of every PUBLICATION issued therefrom, and the exact Number printed thereof.
AT THE CONCLUSION IS GIVEN
A CURIOUS DISSERTATION ON THE
ORIGIN OF THE USE OF PAPER;
Also, a complete HISTORY of the ART of
WOOD-CUTTING AND ENGRAVING ON COPPER,
From its first Invention in ITALY to its latest Improvement
IN GREAT BRITAIN;
concluding with the Adjudication of
LITERARY PROPERTY;
Or the LAWS and TERMS to which Authors, Designers, and Publishers, are separately subject.
With a Catalogue of remarkable BIBLES and COMMON PRAYER-BOOKS, from the infancy of Printing to the present time.
EXTRACTED FROM THE BEST AUTHORITIES,
BY HENRY LEMOINE, BIBLIOP. LOND.
LONDON, 1797:
Printed and Sold by S. FISHER, No. 10, St. John's Lane, Clerkenwell; also sold by Lee and Hurst, No. 32, Paternoster Row.
BY HENRY LEMOINE, BIBLIOP. LOND.
LONDON, 1797:
Printed and Sold by S. FISHER, No. 10, St. John's Lane, Clerkenwell; also sold by Lee and Hurst, No. 32, Paternoster Row.

Figure 5. Shortening Long Title Page Data

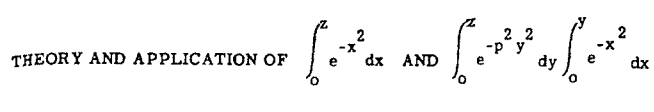

Figure 6. A Mathematical Formula in a Title

But even in modern times there are several ways to write longevity (shown in figure 10): with seven pen strokes (the conventional Japanese form, which is also used as an abbreviated form in Chinese handwriting), with fourteen strokes (the traditional Chinese form), and with fifteen strokes (a Chinese variant that is used symbolically).

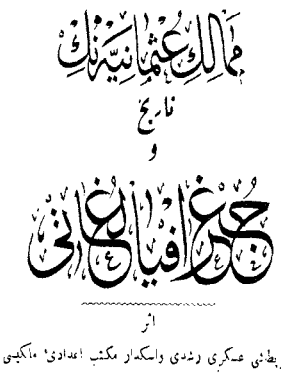

Figure 7. Ottoman Turkish Book on Geography

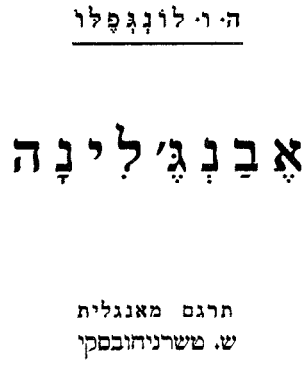

Figure 8. Henry Wadsworth Longfellow's *Evangeline* in Hebrew

If each form was to be uniquely encoded, as some people insist, what is the effect on searching and retrieval? If you are searching for *longevity* as a concept and don't care how it's written, can the system help you? Retrieval based purely on character encoding would require you to enter all the different forms for this kind of search. What if you only knew one of them? What about cross-catalog searching using Z39.50?

This issue is not restricted to ideographs. A library character set, ISO 5426-2 contains a selection of Latin contractions found in early printed works that emulate the manuscript tradition (IOSID 1996). Later editions of the same works do not include contractions. Are added entries or uniform titles the answer to such retrieval conundrums? If added entries are used, how should a result with both forms be sorted? Where does a contraction file relative to the spelled out form?

The British Library character set that was the basis for ISO 5426-2 was appropriate for the technology of its time, to meet the goal of representing the work as accurately as possible in the bibliographic record. The distinction between character and glyph had not been published. At that time, you encoded what you saw and what you needed.

But how much exactitude is needed in a cataloging record today and why? If the body of the entry is intended to be a surrogate for the work, a digital reproduction is a much more faithful representation.

The economic cost of adding variant forms includes: more elaborate input strategies; more complex software to match the variants; more time to create the record if additional access points have to be included; and more complexity in retrieval, for both the interface designer and the user.

So we need to think about our needs for characters at a different level, related not to production of a surrogate for the physical item that has been cataloged, precise in every typographical detail, but to facilitating retrieval. What infor-

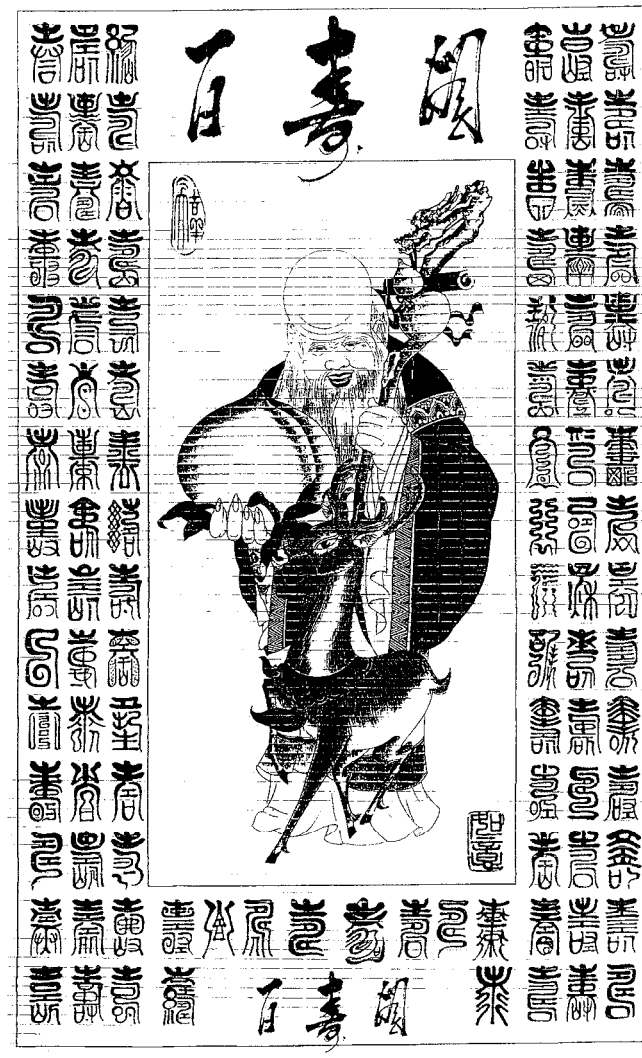

Figure 9. The Hundred Forms of *shòu*

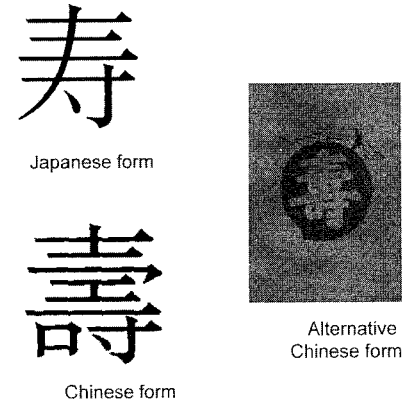

Figure 10. Three Common Ways to Write "Longevity"

mation has to be searchable? Will the user be searching for exact graphic forms, or is a less exact "generic" or "normalized" substitute, supplemented by digital images, sufficient? How do the new text encoding standards (e.g., SGML) relate to traditional cataloging? What sorting order do our users expect? Is it possible to define expectations for the amount of typographic detail that should be captured in a transcribed representation? That is, to define when the digital image will have to be examined? What are the implications for cross-catalog searching when there are different levels of typographic detail in different catalogs?

A cataloging record is generally equivalent to the "plain text" of Unicode. But sometimes it incorporates aspects of rich text: for example, the Latin contractions in ISO 5426-2 or variant ideographs that are differentiated in EACC but represented by a single character in Unicode.

Rather than proposing the addition of ever more characters to Unicode and ISO 10646 to support ever more exact cataloging, we need to define specific needs and identify the best technologies to fulfil them. In particular, we need to examine the potential of SGML and XML, and their relationship to MARC formats.

Needs can be broadly defined as: (1) identification and retrieval, and (2) bibliographic scholarship. The catalog record has to meet the first need. With regard to bibliographic scholarship, a "plain text" catalog record cannot provide all the required detail. A "rich text" form (such as an SGML transcription) or a digital image is called for, and, in some cases, only examination of the actual item will meet the scholar's needs.

I hope this article has answered some of the questions about the Unicode Standard. We need to think long and hard about the characters needed for library catalogs. We need to think about where the balance point is between exact transcription on the one hand and optimized retrieval and sorting on the other. And we need to think about which technologies are appropriate for which user needs.

Works Cited

[All hyperlinks accurate July 24, 1998.]

American National Standards Institute. 1990. *East Asian character code for bibliographic use.* New Brunswick, N.J.: Transaction. (ANSI Z39.64-1989).

Character set and language negotiation [Z39.50 definition]. 1998. http://lcweb.loc.gov/z3950/agency/defns/charsets.html.

CoBRA+ Computerised Bibliographic Record Actions. Factsheet (Feb.). 1998. http://portico.bl.uk/ (click on Information and choose link for CoBRA+ from the page).

Davis, Mark, and Ken Whistler. 1999. Unicode collation algorithm (Draft Unicode technical report #10). www.unicode.org/unicode/reports/tr10/index.html.

Discussion Paper No. 111: Alternate graphics without 880 in Bibliographic, Holdings, Authority, and Community Information Records. 1998. http://lcweb.loc.gov/marc/marbi/dp111.html.

Fisk, Martin, and Anthony Brickell. 1997. *Character set standardisation: Migration strategies to Unicode for National Bibliographic Databases. Final report of the CHASE Project.* PROLIB/CoBRA-CHASE 10169. www.bl.uk/gabriel/cobra/chase.pdf.

Holt, B. P., and Sarah H. McCallum, eds. 1994-. *UNIMARC manual: Bibliographic format*, 2d ed., N.S., vol. 14. Saur, Munich: UBCIM Pub.

IETF Policy on Character Sets and Languages. 1998. Request for comments: 2277. Category: Best current practice (Jan.). http://info.internet.isi.edu:80/in-notes/rfc/files/rfc2277.txt

International Organization for Standardization, Information, and Documentation (ISO/IED). 1996. *Extension of the Latin alphabet coded character set for bibliographic information interchange–Part 2: Latin characters used in minor European languages and obsolete typography.* Genève: ISO 5426-2:1996(E).

International Organization for Standardization/International Electrotechnical Committee (ISO/IED). 1993. *Information technology—Universal multiple-octet coded character set (UCS), Part 1: Architecture and basic multilingual plane.* Geneva: ISO/IEC.

ISO/IEC JTC1/SC22/WG20. 1997. International string ordering—Method for comparing character strings and description of the common template tailorable ordering. ISO/IEC FCD 14651 (Nov. 5). http://osiris.dkuug.dk/JTC1/SC22/WG20/docs/projects#14651.

MARBI Unicode Encoding and Recognition Technical Issues Task Force. 1998. Proposal no. 98-18: Unicode identification and encoding in USMARC records. http://lcweb.loc.gov/marc/marbi/1998/98-18.html.

Network Development and MARC Standards Office, Cataloging Distribution Service, Library of Congress. 1994. *USMARC specifications for record structure, character sets, and exchange media.* Washington, D.C: Library of Congress.

Rather, Lucia. 1968. Special characters and diacritical marks used in roman alphabets. *Library Resources & Technical Services* 12: 285-95.

USMARC to Universal Character Set mappings. 1998. http://lcweb.loc.gov/marc/marc2ucs.html.

World Wide Web Consortium (W3C). 1998a. HTML 4.0 Specification. www.w3.org/TR/REC-html40/.

———. 1998b. Extensible Markup Language (XML) 1.0. W3C Recommendation. www.w3.org/TR/REC-xml.

Unicode Consortium. 1996. The Unicode Standard, version 2.0. Appendix A.2: UTF-8.

Wolf, Misha, et al. 1998. A standard compression scheme for Unicode (Unicode Technical Report #6). www.unicode.org/unicode/reports/tr6.html.