Unicode and Aleph: Questions and Answers for Harvard Library Staff

Q: What is Unicode?
A: Unicode is a character encoding system designed cooperatively by the International Organization for Standardization (ISO) and the Unicode Consortium, a group of hardware and software producers. It is also known as the universal character set (UCS) because of the very large character repertoire it includes. You may also see Unicode referred to as UTF-8. UCS, UTF-8 and Unicode are not technically synonymous, but may be considered so for many purposes.

Unicode is becoming the industry standard, whose development has been propelled by the need of many companies to do business world-wide. Libraries are the beneficiaries of this development. Unicode is the internal character set of Windows, from NT forward; it is supported by recent versions of major web browsers, and by database vendors such as Oracle.

Q: Why do I care?
A: Unicode is also the character encoding used by Aleph. It differs in a number of ways from the character encoding used by HULPR and the first HOLLIS. The changes will affect the way you enter data in Aleph if characters other than the “English” alphabet, numbers, and common punctuation are involved.

Q: Will the adoption of Unicode affect the way OPAC users or I will search in Aleph?
A: No, except in a few rare cases, because index terms and search arguments will both be “normalized” as they have been in HOLLIS since 1985. Diacritics and other modifying marks will be continue to be removed from letters, punctuation will be eliminated using similar rules, and characters not readily input from the standard keyboard will have alternative representations; e.g. Þ will continue to be searchable as th. It will be possible, however, to do searches using CJK characters at work stations equipped with the right software.

There are some small changes in normalization that result from changes in U.S. practice and are not related to Unicode; namely, the ampersand and the plus sign will no longer be removed, but will become characters recognized in index terms and search arguments.

Q: What is the advantage of Unicode over former encodings?
A: The major advantage is in the size of the repertoire. The encoding used by MARC21 from its beginning in the 1960s and approximated by HULPR represents most characters by a sequence of 8 bits. (CJK characters in MARC21 are exceptional, requiring 24 bits each.) Unicode allocates 16 bits per character. This means it is possible to represent over 65,000 characters directly by single values (code points) as opposed to fewer than 256.
Q: How have we gotten along with such a small set?
A: MARC in the United States has used several strategies to deal with the limitation. First, it adopted the practice of entering modifying marks, such as diacritical and tone marks, as separate characters next to their base letters. Later, to accommodate non-Latin scripts—specifically Greek, Hebrew, Arabic, Chinese, Japanese, and Korean—it employed standard data-processing techniques to “escape” to other character sets designed for those scripts. This type of encoding and the character repertoire encompassed by it have recently been named MARC-8. Unicode now allows all the MARC-8 scripts to be encoded directly without need for the cumbersome escape process. Using Unicode, it is also theoretically possible to catalog languages such as Armenian, Thai, or Hindi in the original scripts for the convenience of readers of those languages.

Q: Can I start using all those characters now?
A: No. There are many issues to be considered before expanding the character repertoire of HOLLIS as we change to Unicode. Of great importance among these is adherence to national standards and conventions when communicating bibliographic and holding data. Because Aleph is ahead of most library systems in adopting Unicode, we must initially limit the character repertoire to be used at Harvard to characters that are in the MARC-8 set or that can successfully be converted to characters in that set. This will enable us to export records to other systems, notably OCLC and RLIN, which are not yet ready to accept Unicode records.

There are also local support questions, some trivial, some not, that must at least be investigated before each expansion to the character repertory is approved for HOLLIS use.

Q: How will I know what characters are permitted?
A: OIS will provide a list of permissible characters. Characters not listed must not be used. The names of some characters, especially modifying marks, vary from MARC21 to Unicode. The list will give both names where there are differences. Most HOLLIS documentation will employ the Unicode terminology, often somewhat abbreviated.

Q: What are “characters that can successfully be converted to MARC-8”?
A: In the large Unicode repertoire there are code points assigned for many characters that in MARC-8 must be represented as two or three characters—a base character and one or two modifying marks. Such code points are said to be “precomposed” if there are also Unicode code points that define the parts separately as letters and “combining diacritics.” For example an e with acute accent can be represented in Unicode by one character or by two. In the latter case, the character is said to be “decomposed.” In MARC-8 the accent and the e must be represented by two adjacent characters, because MARC-8 characters are always decomposed. (A display may place the accent properly over the e or it may show it preceding the letter; so, display does not reliably reveal how the character is stored in the database.)

A Unicode character that can be decomposed into parts that are included in the MARC-8 character set can be successfully converted to MARC-8 for export. In general, such characters may be used in Harvard Aleph. Not every precomposed Unicode character meets this criterion because the Unicode repertoire of combining diacritics includes some marks not defined in MARC-8. However, every MARC-8 diacritic is present in the Unicode repertoire.

Q: Which is preferred in Aleph, precomposed or decomposed forms?
A: In general, precomposed forms should be used in Aleph at Harvard. They are generally better supported by browsers, and the tools provided to facilitate data entry are set up to favor them. Another big advantage is that you do not have to be concerned about the proper sequence of the parts of a decomposed character. Use of precomposed forms will reduce the likelihood of input errors. For it is an inconvenient fact that MARC-8 requires the diacritic to precede the base letter while in Unicode the base letter must come first. When you are working in Aleph you must use Unicode conventions and when you are working in OCLC or RLIN, you must use MARC-8 conventions.

Q: What are the exceptions to the use of precomposed forms?
A: There are three types of characters that are to be entered decomposed. First, there are a few precomposed character values fairly new to Unicode that do not display in the font that we are supporting for staff and public catalog interfaces. Second, there are a number of characters used in the LC romanization of certain scripts, notably those used in Indic languages, that do not exist as precomposed Unicode characters. This is natural since the intent of Unicode is to enable languages to be represented in their original scripts. The third type involves the diacritics that span two characters—the ligature used in romanization of Cyrillic and the double tilde found in Tagalog. Processing of these for conversion between Unicode and MARC-8 is facilitated if the diacritics themselves continue to be represented in two pieces as has always been the case in MARC.

There also may be characters in infrequently encountered Latin-script languages for which the input tools do not provide precomposed forms. If you encounter such a character, input it decomposed, but please document its occurrence and report it to OIS for possible addition to the list of authorized precomposed characters.

Q: What are these tools?
A: Different tools are available for people with different needs. If you frequently enter data involving characters not on the standard keyboard, your work station will be equipped with the Macro Express Character File and instructions on how to assign characters to keys or key combinations. If you only occasionally need characters not on the standard keyboard, the Aleph Floating Keyboard, distributed with the client, will enable you to paste these characters into your data. There is also a way, using a function key, to enter the hexadecimal Unicode value of a character. More information on all three of these methods is provided in other documents.

Q: I have seen a web page showing how to enter characters into MS Word using the alt key and a decimal number. Can I use that method in Aleph?
A: No. Do not do that. Not all the characters on that page are valid MARC21 characters, and some of the values supplied do not match Unicode values. Instead, use one of the techniques mentioned above.

Q: Why don’t OCLC and RLIN use Unicode?
A: HOLLIS is a large and complex system, but it is small when compared to either of those utilities. Change some slowly to big systems. OCLC is in the midst of designing a new one; it will use Unicode. There is not a clear timetable for its implementation. RLIN is also aware of the need to change and is currently working on being able to take in Unicode records. Whether or when it will convert its database is not clear. Meanwhile, we must observe the differences between MARC-8 and Unicode as we work in various environments.