

Multilingual Cataloguing: practical and technical issues

Cathie Jilovsky
Information Services Manager
CAVAL Collaborative Solutions

cathiej@caval.edu.au

Andrew Cunningham
e-Diversity and Content
Infrastructure Solutions
Public Libraries Unit, Vicnet
State Library of Victoria

andrewc@vicnet.net.au

Abstract:

Victoria, Australia is a culturally diverse society. The state has a population of four million, of whom one quarter were born overseas. Providing library services in a range of community languages has long been important. The advent of the internet and development of Unicode has provided many opportunities to advance these services.

The first part of the paper will focus on multilingual cataloguing and Unicode issues from a practical perspective. The complexities of implementing and operating a system which supports a multi-language environment are explored, using the implementation of the Ex Libris Aleph 500 library system at CAVAL Collaborative Solutions as a case study. CAVAL provides cataloguing services to customer libraries in over sixty languages. CAVAL, a consortium of the University Libraries in Victoria and the State Library of Victoria, provides a range of services to the library and information communities throughout Australasia.

The second half of the paper will describe technical issues that relate to multilingual cataloguing. These include Unicode and software, web and xml internationalization which impact on library websites and library systems, with a focus on non-roman alphabets. Examples will be drawn from the experience of providing multilingual computing facilities and a multilingual portal for public libraries through the State Library of Victoria.

Introduction

Victoria, Australia is a culturally diverse society. The state has a population of four million, of whom one quarter were born overseas. Providing library services in a range of community languages has long been recognised as important. The advent of the internet and development of Unicode has provided many opportunities to advance these services.

Unicode is an international character set developed for the scripts of all the languages of the world. It is becoming a fundamental tool for communicating and exchanging data between computer systems. The power of Unicode is that it enables human communication across languages, cultures and geographic areas. However, although it is complex to understand and to implement, it provides for system interoperability that is essential in the electronic age. Unicode must be embraced by libraries and information services in order to remain relevant in the wider community.

Unicode was devised so that one unique code is used to represent each character, even if that character is used in multiple languages. It is a complex standard and is not possible to describe without some technical detail. The major computer companies were the drivers in the development of Unicode, and its incorporation into many commonly used software products today provides a satisfactory platform for the support of multiple languages. [Wells 2000]

CAVAL Collaborative Solutions, a consortium of the University Libraries in Victoria and the State Library of Victoria, provides a range of services to the library and information communities throughout Australasia. One of these services is the provision of multi-language cataloguing. As a result, when CAVAL staff surveyed the library system marketplace in 2001, the ability to use the Unicode format was a key factor in the selection of a system. The system selected and subsequently purchased was the Aleph 500 system from Ex Libris.

The implementation process proved to be quite complex. The Implementation Team had to grapple with a range of issues - ranging from understanding the MARC21 and Unicode standards; to developing internal workflows and procedures as well as processes for the transfer and conversion of records to and from a variety of external systems. At this time many of the external systems either did not support Unicode at all or only supported a portion of it. Now, in 2005, more library systems are starting to use Unicode, however it is still not commonplace, at least in Australia. This presented the team with a range of challenges to be addressed. In particular, reference is made to the National Library of Australia's Kinetica¹ system, as like most Australian libraries, CAVAL sources from and contributes high quality records to the National Bibliographic Database.

Usage of Unicode in libraries

A review of the library literature identified a number of papers discussing and describing the Unicode standard and others seeing it as a component of the growing digital library environment, but very few with details of library-specific

implementations. Issues identified included the incorporation of Unicode into library systems, the affect of Unicode on the exchange of records, language translation tools, sorting mechanisms, dealing with word parsing in specific languages and developing character mapping tools. Tull's report on a survey of library system vendors found that that most companies do not plan to support all the scripts in the world, but are concentrating on the ones that their customers need. [Tull 2002]

Library system vendors are now incorporating Unicode into their products. A number of these systems can store data internally in Unicode, but the literature indicates that the exchanging of library data between systems in Unicode is minimal as yet. Changes will be required to the large bibliographic databases from which libraries source their records in order to support both the storage and exchange of records in Unicode. Data storage requirements vary according to the encoding form, for example if UTF-8 is used, the space requirements relate directly to the character content of the data.

The MARC21 specification specifies the use of Unicode for the exchange of records. Currently the encoding of MARC records is limited to UTF-8, as it is recognised that there will be a period of transition to the 16-bit Unicode environment. Several positions in the MARC record leader take on new meanings e.g. the record length contained in Leader positions 0-4 is a count of the number of octets in the record, not characters. Diacritical marks must be encoded following the base letter they modify which is the opposite of the MARC-8 rule for encoding order. [MARC 21 Specification 2005]

Aliprand points out that in pre-computer catalogue days compromises were often made between a totally accurate transcription and a form which could be easily and reliably found by searchers. This applies with equal validity in the automated environment, i.e. facilitating retrieval does not necessarily mean being precise in every typographical detail. [Aliprand 2000]

Diacritics (or accent marks) are used to remedy the shortcomings of the ordinary Latin alphabet for recording languages which have additional sounds. [Wells 2000] Transliteration is the spelling or the representation of characters and words in one alphabet by using another alphabet. In the English speaking world, romanisation is commonly used in library catalogues. However to achieve this, specialised knowledge is required by both library cataloguers and users. Other libraries have taken a simpler approach and ignored diacritical distinctions. [Erickson 1997]

The multi-language environment at CAVAL

CAVAL's cataloguing staff can catalogue in over 60 languages. As Australia's pre-eminent provider of multi-language cataloguing, CAVAL has a pool of professional, experienced language specialists and is always happy to seek out skills in additional languages. In partnership with a Melbourne book supplier, the Foreign Language Bookshop (see <http://www.flb.com.au/default.asp/>), shelf-ready foreign language material, including ESL (English as a second language), travel guides, audio/video, and preselected packs of popular titles for libraries can be selected, catalogued and

processed. Assistance with the selection of materials and/or translation of publishers' brochures is also provided.

Languages in which CAVAL cataloguers are currently working are: Afrikaans, Albanian, Arabic, Armenian, Bengali, Bosnian, Breton, Bulgarian, Burmese, Chinese (Pinyin and Wade-Giles, traditional and simplified script), Croatian, Czech, Danish, Dari, Dutch, Esperanto, Farsi, Finnish, French, German, Greek, Gujarati, Hebrew (Classical and Modern), Hindi, Hungarian, Indonesian, Italian, Japanese, Jawa, Khmer, Korean, Kriol, Kurdish, Latin, Latvian, Lithuanian, Macedonian, Malay, Maltese, Norwegian, Portuguese, Polish, Portuguese, Punjabi, Romanian, Romansch, Russian, Serbian, Sinhala, Slovak, Slovenian, Somali, Spanish, Swahili, Swedish, Tagalog, Tamil, Tetum, Thai, Turkish, Ukrainian, Urdu, Vietnamese, Welsh, Yiddish, Yoruba, Zulu. In addition to cataloguing, services provided to customers include translation, transliteration and abstracting.

Implementing the Aleph system at CAVAL

CAVAL uses the Aleph system for two separate applications. The first is to support the CARM Centre collection and the second is for the cataloguing services that CAVAL provides for external customers. The CARM (CAVAL Archival and Research Materials) Centre is a high-density storage facility for low-use research materials, and in addition provides space for short-term storage of collections. It has a capacity of 1 million volumes, and provides a humidity and temperature controlled environment.

The first practical issue faced was understanding the terminology associated with Unicode - for example, the distinctions between a character and a glyph, and between fonts and character sets, as well as the details of the MARC21, UTF and Unicode standards. The implementation of software that supports multiple languages entails the consideration of a range of issues beyond the data being stored in Unicode, such as sorting and display. Sorting data containing more than one language is not straight forward, as the sorting order of characters varies between languages. Understanding why a system is unable to display a particular character is often complex, e.g. it may have no information about that character, or may lack the right tool (such as a font) to display that character.

The system was installed in early 2002 and live usage of Aleph for the CARM Centre commenced in May of that year. Considerable effort was required to implement the use of diacritics and subsequent development of workflows and procedures and CAVAL cataloguing staff began using the system for customers in December 2002. A number of other essential processes were implemented and finetuned during the early part of 2003, which included the collection of statistical data, the export of records to Kinetica and the import of records from CARM member systems.

Since that time the team of CAVAL Cataloguers have become proficient users of Aleph and many of them catalogue in two or three or more languages. Procedures and workflows have been streamlined as much as possible and working in a Unicode environment has become a reality. A number of further improvements will be

implemented following the upgrade to Aleph Version 16.02, scheduled for late June 2005.

The Aleph Implementation Team approached the process with enthusiasm but soon discovered that although the theory of Unicode is simple, in practice the implementation is complex. Many hours of consultation, research, discussion and testing were needed before it was felt that sufficient understanding was achieved to use diacritics in Aleph with confidence. After initial training and system analysis, four major areas relating to the use of diacritics in Aleph were identified:

1. Storage and conversion issues

As most of the systems with which CAVAL exchanges records are not yet able to accept data in Unicode format, data encoding conversion routines must be used when exporting and importing records. It took the team some time to understand all the issues relating to the conversion of diacritics in data exported to, and imported from, external systems. Once these were understood a structured testing process was developed. Initially the focus was on transferring data to and from the National Library of Australia's Kinetica system. Staff at the National Library of Australia were extremely helpful during the testing phases and worked closely with CAVAL staff to achieve the best outcomes. The understanding gained from developing the processes for exchanging records with Kinetica were applied to developing the exchange of records with other systems.

2. Data input issues

The initial challenge was to analyse and understand the issues. Once this point was reached it was then a matter of producing documentation and providing training for the cataloguing staff. As CAVAL's cataloguers were all experienced with Kinetica, this encompassed articulating the differences between Aleph and Kinetica client input conventions e.g. the input order of diacritics when combining with other characters is different for Aleph and for Kinetica. It is hoped that in the future as Kinetica moves towards Unicode implementation this will become less of an issue.

3. Display of data containing diacritics

The troubleshooting process again proved to be a complex one. Diagnosis of a problem could potentially involve the operating system, the web browser, fonts used for screen display, fonts used for printing, character encoding and/or the input method editor. It also proved necessary to upgrade the PC environment and to ensure that sufficient technical support was available during the implementation process. It is essential that the Aleph Client software is configured correctly on each PC, that appropriate fonts are installed for each language and that all operating systems, web browsers and printers being used for Aleph are Unicode-aware. Training and documentation was developed by CAVAL staff and provided for the cataloguers.

4. Cataloguing issues

CAVAL's cataloguers are a multi-cultural and multi-lingual team. When recruiting staff selection criteria include language as well as cataloguing skills. An understanding of the associated cultural environment is an important component of cataloguing in other languages. This includes history, geography and politics as well as some knowledge of publishing traditions. The establishment of name headings

can be problematic e.g. the heading for a Russian author who publishes in Australia would have no diacritics, whereas another work published by the same author in Russian would incorporate full diacritics.

Practical lessons

The system implementation was undertaken by a team of CAVAL staff with a range of skills and experience. Once the importance of separating input, storage and display issues was understood it became easier to make practical decisions to resolve problems. These ranged from technical issues such as the configuration of workstations, including operating systems, input methods, web browsers and fonts; to communication issues such as training staff in Unicode concepts and validating screen displays for each language. The language knowledge of CAVAL's multi-lingual and multi-cultural cataloguing staff, the technical skills of the IT staff, a multi-lingual Systems Librarian and staff who all communicated well were essential components of the process.

At this point in time CAVAL is creating Unicode records in the Aleph database, however files of records exported to many customers have the diacritics stripped out. We look forward to working with our customers to improve this situation as their systems are upgraded to be Unicode compliant.

Records created in CAVAL's Aleph system currently contain only romanised bibliographic data. Future developments will include the incorporation of native vernacular scripts. We expect to work with other libraries as they implement Unicode compliant systems, the goal being to import and export Unicode records. However, we will build on the lessons we have learnt during this initial implementation and will plan carefully for a phased approach.

Multi-lingual cataloguing: Technical Issues

Unicode support

Unicode implementations in library management systems (LMS) are shaped by archaic character encoding conventions underlying the MARC standards. Although vendors have started making the shift from character encodings that have their roots in the 1960s towards Unicode, implementations suffer from major internationalization flaws. LMS Unicode development has been driven by the need to support limited numbers of key languages (European and East Asian) and the necessity to support archaic transcription conventions.

What do Vendors mean when they refer to Unicode support? Essentially Unicode support can be summed up as:

- Do not use unassigned codepoints
- The implementation can be restricted to a Unicode subset
- Software needs to take Canonical equivalence into account
- Ignore illegal encodings

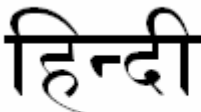

- Right-to-left scripts have to be processed according to the Unicode bi-directional algorithm

Different library management systems could be using different versions of the Unicode standard. The current version is Unicode 4.1. For example, Endeavor Information Systems' Voyager with Unicode release is based on Unicode 3.2. On Voyager data is entered as UTF-8.

One critical element of library management systems is the question of canonical equivalence and normalization. For instance, Microsoft's Vietnamese keyboard layout uses unique pre-composed characters for Vietnamese vowels and uses combining diacritics for tone markers. Alternatively, most third party Vietnamese input software uses fully pre-composed characters, i.e. single unique characters. It is possible to fully decompose characters forming a base character and one or more combining diacritics. All of these forms are equivalent and should be treated the same.

tiếng Việt (NFC)	tieôóng Vieot (NFD)	tiông Viêt (Microsoft)
----------------------------	-------------------------------	----------------------------------

Normalization is the process of converting text to either full pre-composed character sequences or fully decomposed character sequences. Normalization is crucial in order to ensure matching of search patterns.

	
---	--

The additional factor that needs to be taken into account is that some writing scripts require complex processing. Not all scripts have a one to one relationship between character and glyph. Some writing scripts reorder the position of the glyph; some change the shape of the glyph and may move it above, below or in front of the base character. The Hindi sample above shows the fully rendered version of the word on the left, and the sequence of Unicode characters that form this work on the right.

The capability of the software to render complex scripts is dependant on the operating system. Windows 2000, Windows XP and Windows XP Service Pack 2 contain increasing support for additional character scripts.

Behind the scenes

A library management system should be able to:

- Store and manipulate Unicode strings
- Handle conversion to and from other character encodings including MARC8 and EACC

- Handle formatting/parsing functionality for numbers, currencies, date/time and messages for all locales needed
- Handle Unicode-conformant collation, normalization, and text boundary (grapheme, word, line-break) algorithms
- Use Charset-independent locales (all Unicode characters usable in any locale)

In addition windows clients should be able to:

- Display, print and edit Unicode text
- Handle BIDI display if right-to-left characters are supported, e.g. Arabic, Hebrew, Syriac, etc
- Handle character shaping if scripts such as Arabic and Indic are supported. On the windows operating system the client should be able to utilise Microsoft's Unicode Script Processor (Uniscribe).
- Input text using Input Method Editors (e.g. Chinese, Japanese and Korean input methods)
- Be able to select appropriate fonts for editing, and make use of font linking technologies in order to use appropriate fonts for each writing system supported. This should be customisable by the end user.

The Voyager client allows the selection of an appropriate font for cataloguing. It is possible to change fonts when the language being with changes. The key limitation is that it requires the use of a font that supports both the basic Latin script and the language you wish to catalogue in. Many of the new scripts supported on Windows 2000 and Windows XP are using fonts that do not contain the basic Latin characters, and so would be unsuitable for the Voyager cataloguing client.

Web interface

The weakest link in most library management systems is the web interface to the catalogue. The World Wide Web Consortium's (W3C's) Internationalization Guidelines, Education and Outreach working group has been developing web internationalization authoring guidelines. These guidelines indicate appropriate methods to:

- Identify the character encoding of a web page;
- Identify the language of a web page, or indicate change of language; and
- Appropriate methods to mark up web pages with bi-directional text.

Language tagging is not only required for web internationalization. The WAI accessibility guidelines also require change of language to be indicated.

The websites of the State Library of Victoria and Victorian public libraries belong in the vic.gov.au domain and as with other websites within this domain are required to adhere to state government web standards. The state government's accessibility standard requires level A compliance. In other words, any change of language needs to be indicated in the XHTML/HTML markup. The current generation of library catalogue web interfaces do not include language tagging. This provides problems from both the point of view of web accessibility and of web internationalization.

In some instances, language tagging is crucial. It is necessary for Unicode CJKV

text. In Unicode, many hanzi, kanji and hanja have been unified and use the same codepoint. There is often variation between the glyphs used for Simplified Chinese, Traditional Chinese, Japanese and Korean and separate fonts are used for each language using culturally appropriate glyphs.

Modern web browsers use language tags to select appropriate CJK fonts in the absence of any other mechanism indicating which font to use. If the language of a web page has the value of “zh-CN” a simplified Chinese font will be used if the web page author did not specify a font. The browser will use a traditional Chinese font for web pages that are marked as “zh-TW”, a Japanese font for text tagged as “ja” and a Korean font for text marked up as “ko”.

Problems begin to arise when web pages using Han ideographs do not have the language indicated. In the absence of language tags, Internet Explorer and Mozilla will default to using a Japanese font for all CJK text in the absence of any indication of the language of the text. Opera will default to a Korean font.

Most library management systems do not indicate language. Additionally many vendors recommend that libraries should use large multi-script fonts to display their web based catalogues. The Voyager documentation recommends the use of the font Arial Unicode MS. Other vendors also suggest fonts such as Andale Mono. There are a number of problems with this approach. Firstly fonts are restricted to a maximum number of characters. It is not possible to create a font that encompasses the whole Unicode character repertoire. There is also the issue of using appropriate glyphs for Han ideographs. The font Arial Unicode MS, which is most often recommended, is a commercial font that shifts with various Microsoft Office applications. Installing this font on a computer that doesn't have an appropriate Microsoft Office application installed would be a license violation. At least two versions of Arial Unicode MS have been released. The older version does not appropriate support for the Devanagari and Tamil scripts, while the later one does.

Libraries require much more precise control over the fonts used to display non-Latin script records. It is also critical to remember that the ability of Internet Explorer or Mozilla to display records in particular writing scripts, isn't so much based on the abilities and limitations of the web browser, rather it is defined by the abilities and limitations of the operating system.

Conclusion

CAVAL's requirement to catalogue in a large number of languages, both European and Asian, provided particular challenges in the implementation of a Unicode-compliant library management system. However following careful analysis of the issues, a phased approach was adopted and the system was successfully implemented. The staff involved developed considerable Unicode expertise including skills in system configuration and the development of suitable workflows, procedures and staff training. Collaboration with staff at the State Library of Victoria who are also working with Unicode systems has been invaluable.

A high degree of system interoperability is now possible with the integration of

Unicode into many components of the computer environment. Library system vendors must be involved in the further development of software and user interfaces and work with libraries to enable true multi-lingual access for the global community.

References

Aliprand, J.M. 2000. The Unicode standard: its scope, design principles and prospects for international cataloguing. *Library Resources and Technical Services*, vol. 44, no. 3, pp. 160-167.

Cunningham, A. 2004. Global and local dimensions of emerging community languages support. VALA2004 12th Biennial Conference and Exhibition, Breaking Boundaries: Integration and Interoperability, Melbourne Australia. <http://www.vala.org.au/vala2004/2004pdfs/35Cuning.PDF> Accessed 16 Jun 2005.

Erickson, J.C. 1997. Options for presentation of multi-lingual text: Use of the Unicode standard. *Library Hi Tech*, vol. 15, no. 3-4, pp. 172-188.

Jilovsky, C. 2004. Unicode: a tool for system interoperability and human communication. VALA2004 12th Biennial Conference and Exhibition, Breaking Boundaries: Integration and Interoperability, Melbourne Australia. <http://www.vala.org.au/vala2004/2004pdfs/45Jilov.PDF> Accessed 16 Jun 2005.

MARC 21 Specification. <http://www.loc.gov/marc/specifications/spechome.html> Accessed 16 June 2005.

Tull, L. 2002. Library systems and Unicode: A review of the current state of development. *Information Technology and Libraries*, vol. 21, no. 4, pp. 181- 185.

The Unicode Standard: A Technical introduction, 2005. <http://www.unicode.org/standard/principles.html>. Accessed 16 Jun 2005.

There is extensive documentation of fonts, code pages and character sets on the Microsoft web site <http://www.microsoft.com/typography>.

Endnote

ⁱ Information about Kinetica and the Australian National Bibliographic Database can be found at <<http://www.nla.gov.au/kinetica/aboutkinetica.html>>