Multi-lingual Cataloguing: culture, practice and systems

Cathie Jilovsky, Lamis Sukkar, Eva Varga

CAVAL Collaborative Solutions

Abstract:
The provision of multi-lingual web services is dependent on the creation and maintenance of appropriate multi-lingual metadata. In particular multi-lingual cataloguing data is an essential component of providing access to digital libraries and library catalogues.

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, currently CAVAL provides cataloguing services to customer libraries in over sixty languages. Customer libraries come from all sectors - public libraries, state libraries, academic libraries and government and special libraries.

Unicode is an international character set developed for the scripts of all the languages of the world. The ability to use the Unicode format was a key factor in the selection of a library management system to support CAVAL's multi-lingual cataloguing services. The Aleph 500 library management system from Ex Libris was implemented in 2002, and the system was upgraded in 2005. The CAVAL staff involved in the implementation and operation of this system to meet the evolving needs of the CARM (CAVAL Archival and Research Materials) Centre and the CAVAL Cataloguing business have developed considerable expertise in the use of Unicode. A range of issues, such as understanding the MARC21 and Unicode standards, developing internal workflows and procedures, developing processes for the transfer and conversion of records to and from a variety of external systems, have been grappled with and solutions developed.

CAVAL's cataloguers are a multi-cultural and multi-lingual team. Many staff have skills in several languages. In addition to traditional library cataloguing, multi-lingual services provided to customers include translation and transliteration, metadata, web sites and abstracting. When recruiting staff selection criteria include language as well as cataloguing skills. We believe that an understanding of the associated cultural environment is an important component of cataloguing in other languages.

This paper will explore three aspects of multi-lingual cataloguing, using CAVAL as a case study

- *Culture: the mix of cataloguing skills and cultural and language knowledge needed by staff*
- *Practice: the development of appropriate procedures and processes*
- *Systems: the complexities of implementing and operating a system that supports a multi-language environment.*

Introduction

The provision of multi-lingual web services is dependent on the creation and maintenance of appropriate multi-lingual metadata. Metadata is often defined as 'data about data', it is the information that enables searchers to locate resources, whether the resources are web pages, or links to other types of resources, for example print materials. Library catalogues contain metadata in the form of records consisting of bibliographic data that describe library resources. Today's library catalogues provide access to both physical and digital resources. The digital resources are likely to be a mix of freely-available web resources and others which are available through paid subscriptions. An essential component of the provision of library services for multi-lingual communities is the availability of multi-lingual cataloguing data.

CAVAL Collaborative Solutions, a consortium owned by a number of Australian universities, provides services to its members as well as being a long-standing provider of services to the information and knowledge sectors throughout Australasia. These services include the provision of multi-language cataloguing and working with libraries in their implementation, management and operation of library systems. Currently CAVAL provides cataloguing services to customer libraries in over sixty languages. These customer libraries come from all sectors - public libraries, state libraries, academic libraries and government and special libraries.

CAVAL's cataloguers are a multi-cultural and multi-lingual team. Many staff have skills in several languages. In addition to traditional library cataloguing, multi-lingual services provided to customers include translation and transliteration, metadata, website services and abstracting. When recruiting staff selection criteria include language as well as cataloguing skills. CAVAL believes that an understanding of the associated cultural and historical environment is an important component of cataloguing in the languages used within and by that culture or geographic area.

The advent of the internet and the development of Unicode has provided many opportunities to advance library services, especially in the multi-lingual environment. 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.

This paper will explore three aspects of multi-lingual cataloguing - culture, practice and systems using CAVAL as a case study

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. [Henczel and Monester 2002]

In addition to cataloguing, multi-lingual services provided by CAVAL include translation, transliteration and abstracting.

Multi-lingual cataloguing: Culture

CAVAL's cataloguers are a multi-cultural and multi-lingual team. Cataloguers of foreign language materials need to possess sound linguistic knowledge, skills and abilities and be familiar with the general culture of the language and the geographical region, as well as historical and political influences. High-level formal education within the culture is also desirable.

There are a number of significant factors that may have a profound effect on foreign language cataloguing practices. These include political changes and historic events. For example recent political changes in the former Soviet Union and Eastern Europe brought with them changes in publishing practices, revisions in Library of Congress practices, and even changes in official scripts used in some of the new political entities. With the creation of new countries changes to authorities such corporate bodies and geographic names are needed. Many of these changes involve different language forms, while others demand historical research. [Princeton 1998]

The establishment of name headings can be problematic. For example the heading for a Russian author who publishes in Australia would have no diacritics, whereas another work published by the same author in Russia would incorporate full diacritics. Determining the language and even the script of the material in hand is not always straightforward. For example before the Yugoslav wars, Serbo-Croatian or Croato-Serbian was used in Bosnia, Croatia and Serbia, with some geographic pronunciation variations, and Cyrillic script was used in Serbia. Now there are three "separate" languages.

Several languages may use the same script. For example Arabic script is used by about 20 languages including Kurdish, Persian and Urdu, with some variations in orthography. Other languages use several scripts. The cataloguing description must indicate when an item is in a script other than the primary one for the language, for example the primary script for Azerbaijani is Cyrillic, but it can also be Arabic or Roman.

What name is it?
Following the collapse of the Soviet Union, each new republic gained a new official language. Consequently, in areas other than Russia (Federation), the headings of all cities must be in the language of the country (e.g. Armenian cities in Armenian, Uzbek names in Uzbek). This means that practically every heading must be changed. There are no hard and fast rules for this, and the Library of Congress Name Authorities do not provide consistent examples. [Princeton 1998]

The foreign language cataloguer may be faced with another complication stemming from ambiguities in grammatical forms. At times, it is unclear what a person's name is due to inflected forms. Name headings with post-positional articles (e.g. Bulgarian names) must be based on usage, therefore some detective work may be required.

Similarly, when the surname is clearly not in the author's native language, the cataloguer must attempt to establish the name in the correct form.

The imprint

Incomplete publishing data, inconsistencies in the layout and location of publishing information within the item, incorrect or invalid ISBN numbers are more commonly encountered by the foreign language cataloguer. Many countries do not have strong copyright legislation. Translations may be published without the acknowledgement of the original author and title of the work. The word for printing in many foreign language publications is not used as narrowly as it is in Western publishing, and can be taken to mean publication. Non-English language publishers often use the word 'edition' for an additional printing of the same manifestation. [Princeton 2000]

When the title page bears the title written in elaborate calligraphy, as is often the case with Arabic script, the cataloguer has to look for the title written elsewhere in a simpler design (e.g. on the cover or the spine). Diacritical marks may be absent in printed texts, depending on the style of the book and its typography. For most languages descriptive cataloguing requires a lower degree of language proficiency than subject analysis and classification. However for other languages, such as Arabic and Hebrew, in which most vowels are not written but understood by the reader according to the grammar and context of the word, cataloguers require an in-depth knowledge of the language. Romanisation is not the same as transliteration, the latter implies "simple letter-by-letter substitution." (Vernon (1996) cited by Wilson)

Cultural sensitivities

Personal names in many cultures (African, Arabic, Muslim) can be composed of many different elements. Inaccurate cataloguing of personal names could deeply offend someone from the culture in question. [Bertelsen 1996]

The Library of Congress Subject Headings and classification system reflect a Western bias and are often inadequate for certain materials. Some of the subject headings may even be pejorative and incomplete in the context of a specific culture. (Bertelsen 1996) Many English-speaking and non-English-speaking countries have adopted the Anglo-American Cataloguing Rules; however, problems were encountered by various countries attempting to adapt it to a non-Anglo-American environment or translate it for a non-English-speaking environment. (Stern 1996)

Given the complexities as shown in the examples above, cataloguing of non-English materials may often require more time to be spent in consultation of reference works, searching the rules for guidance, or maintenance of the catalogue. The guidelines provided by the Library of Congress are often late in coming, inconsistent or inadequate. Impact of these issues on cataloguing productivity has to be therefore carefully considered.

Multi-lingual cataloguing: Practice

The practical aspects of managing the work practices for CAVAL Cataloguing staff working in more than 60 languages for about 45 customer libraries are complex. This mix of languages and scripts has a significant impact on cataloguing practices and has necessitated the development and maintenance of special and unique

procedures and processes. High standard practices include the use of *ALA-LC Romanisation Tables: Transliteration Schemes for Non-Roman Scripts*, which is also used by the National Library of Australia, and ensuring that the correct diacritics are input into cataloguing records. The ALA-LC Romanisation tables cover more than 150 non-roman languages. [ALA-LC 2004]

The two main issues that CAVAL considers when specifying and developing procedures for multi-lingual cataloguing projects are

1. The customer's library and library management system.
2. Cataloguers and their expertise. Do we have all the cataloguers with the language skills needed to complete the project successfully?

The Library

The challenge of creating high standard records for the customer and using the appropriate Library of Congress transliteration tables and thus the correct Unicode characters is always the first priority to be considered. Questions asked will include

- What level of cataloguing standard does the customer require? Will the use of minimal level standards be sufficient? Or is cataloguing to high level standard, defined as AACR2 cataloguing level 3 required? The minimum standard is -
  (i) AACR2 level 1 description
  (ii) A minimum of one name or uniform heading if applicable
  (iii) Include the following fields: Title proper, statement of responsibility, Publication details and Imprint information
- What are the capabilities of the library management system? Can diacritics be entered into each field?
- Can the library management system display and print the diacritics?
- Can the library management system display the vernacular script if input in the relative relational fields?
- What method will be used to load and catalogue the records onto the customer's library management system?
  (i) Is the Australian National Database (Libraries Australia) being used for cataloguing?
  (ii) Does the customer require CAVAL to connect to their database and create records directly onto their in-house system?
  (iii) Should CAVAL's own Aleph system be used to create records?

CAVAL works closely with each individual library to ensure their needs are met, and also with the National Library of Australia. Therefore when cataloguing using the National Library Database care is taken to adapt and work with the National Library's own practices. These include the use of Library of Congress standards, and the input of diacritics as stated in their procedures manual.

Cataloguers

Cataloguers are our major asset. CAVAL supports them by

- Providing access to appropriate documentation. These include the Library of Congress transliteration tables and all cataloguing rules in general. New publications are made available in a timely manner, and designated cataloguers summarise current changes and updates to the cataloguing rules on a quarterly basis.

- Appointing a designated cataloguer to be the main contact with each customer library. This cataloguer also has responsibility for creating procedures and specifications, and for training other cataloguers who will be involved in the project.
- Employing a pool of linguistics experts. These experts are essential to the multi-lingual cataloguing operations and have detailed knowledge of Unicode and MARC-8 conversions.
- Running regular in-house training. Sessions may be to enhance existing knowledge, to expand skills with different methods of inputting diacritics, to introduce a new library system or a refresher course.
- Recruiting native speakers to assist with languages for which a cataloguer is not available. The native speaker will transliterate the access points on a special sheet and the data will then be input by a qualified cataloguer.

CAVAL prides itself on working with customers who understand that cataloguing multi-lingual material is a professional endeavour and not purely a data entry operation.

Multi-lingual cataloguing: Systems

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 complex implementation process necessitated the understanding of the MARC21 and Unicode standards, the development of internal workflows and procedures as well as processes for the transfer and conversion of records to and from a variety of external systems. At the time many of the external systems either did not support Unicode at all or only supported a portion of it. Now, in 2006, more library systems are starting to use Unicode, however it is still not commonplace, at least in Australia.

The incorporation of Unicode into library management systems by library system vendors requires a significant shift from the character encoding conventions underlying the MARC standards that have their roots in the 1960s. The development has been driven by the need to support limited numbers of key languages (European and East Asian) and the necessity to support existing transcription conventions. [Jilovsky and Cunningham 2005] The exchange 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. [Tull 2002]

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. For example 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] Diacritics (or accent marks) are used to remedy the shortcomings of the ordinary Latin alphabet for recording languages that 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]

Library systems commonly provide a facility to define words that are automatically ignored when entered at the beginning of a search term. However doing this for initial articles in foreign language titles may create as many problems as it fixes, as words that are articles in some languages are not in others.

Operating 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.

An important practical issue faced during the implementation phase 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. Other issues related to data storage, sorting and display. Sorting data containing more than one language is not straightforward, 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. [Aliprand 2000]

The system was installed at CAVAL in early 2002 and live usage of Aleph commenced in May of that year. Considerable effort was required to implement the use of diacritics and subsequent development of workflows and procedures which included the collection of statistical data, the export of records to Kinetica (now Libraries Australia) and customer systems, and the import of records from CARM member systems. Aspects relating to the use of diacritics in Aleph which were addressed during the implementation phase were storage and conversion issues, data input issues and display of data containing diacritics. The solutions encompassed technical ones such as the configuration of workstations, including operating systems, input methods, web browsers and fonts; and communication ones 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.

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 are now being implemented following the upgrade to Aleph Version 16.02 in 2005.

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 the initial implementation and subsequent operational experience and will plan carefully for a phased approach.

Conclusion

There are many aspects to multi-lingual cataloguing, and there is a range of information available on the web and published in the library literature. This paper has merely described some of the practical issues that CAVAL has experienced and addressed in providing multi-lingual cataloguing services. In particular examples of the range of cultural, historical and political knowledge that staff need along with general cataloguing and specialist language skills have been discussed.

CAVAL's requirement to catalogue in a large number of languages, both European and Asian, provides particular challenges in the operation of a Unicode-compliant library management system. Although Unicode is now integrated into many components of the computer environment, further development of software and user interfaces for library systems is needed. Library system vendors and libraries must work together to enable true multi-lingual access for the global community.

The creation and maintenance of appropriate multi-lingual metadata and multi-lingual cataloguing data are essential components of providing access to digital libraries, library catalogues and multi-lingual web services.

References

ALA-LC Romanisation Tables: Transliteration Schemes for Non-Roman Scripts. 2004. http://www.loc.gov/catdir/cpso/roman.html or http://www.loc.gov/cds/ALARomanization-TOC.html Accessed 9 Jan 2006.

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.

Bertelsen, Cynthia D. 1996. Issues in cataloguing non-Western materials: special problems with African language materials. http://filebox.vt.edu/users/bertel/africana.html Accessed 16 Jan 2006.

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 9 Jan 2006.

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.

Henczel, S. and Monester, A. 2002. Cultivating Non-English Collections: a unique partnership that alleviates the pain of libraries in multi-language communities. American Library Association. http://www.ala.org/ala/iro/iroactivities/cultivatingnonenglish.htm Accessed 9 Jan 2006.

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 9 Jan 2006.

Jilovsky, C. and Cunningham, A. 2005. Multilingual Cataloguing: practical and technical issues. The Multicultural Library: Staff Competence for Success, IFLA Satellite Conference, August 2005, Stockholm, Sweden. http://www.ifla-stockholm2005.se/pdf/IFLA%20Multicultu%C9005%20CJ%20&%20AC.pdf Accessed 9 Jan 2006.

MARC 21 Specification. http://www.loc.gov/marc/specifications/spechome.html Accessed 9 Jan 2006.

Recent trends in Slavic cataloging, Princeton University. 1998. http://infoshare1.princeton.edu/katmandu/sgman/slavchat.html Accessed 9 Jan 2006.

Princeton University Library's cataloging documentation. 2000. Slavic cataloging manual. http://infoshare1.princeton.edu/katmandu/sgman/smtocs.html Accessed 9 Jan 2006.

Stern, B. 1996. Internationalising the rules in AACR: adopting and translating AACR

for use in non-Anglo-American and non-English-speaking cataloguing environments / Barbara Stern. *Cataloging & Classification Quarterly*, vol. 21, no. 3/4.

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 9 Jan 2006.

Wilson, K.E. 2005. A guide to copy cataloging Arabic materials: a Master's paper submitted to the faculty of the School of Information and Library Science of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Science in Library Science.
http://etd.ils.unc.edu/dspace/bitstream/1901/190/1/kristenwilson.pdf Accessed 16 Jan 2006.

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