

Running head: MULTILINGUAL CATALOGING AND ACCESS

An Annotated Bibliography of Resources Related to
Multilingual Library Cataloging and Access

Michael Braun Hamilton

LI804 – Theory of Organization of Information

Emporia State University

**An Annotated Bibliography of Resources Related to
Multilingual Library Cataloging and Access**

This bibliography presents an overview of the history and current developments in the area of multilingual and multiscript library classification. The primary focus is on cataloging and bibliographic control of multilingual materials. The articles are directed primarily towards catalogers and library professionals.

Language issues are becoming increasingly important in the library community. As the cultural and linguistic diversity of our patrons grows, the need for multilingual access becomes increasingly urgent, and the increasing globalization of our profession necessitates improved standards and systems for multilingual bibliographic control. We have come a long way in a short time, and hopefully this bibliography will help highlight some of the outstanding recent accomplishments in this field, as well as lead interested readers toward some of the challenges that still exist towards ensuring equal access to library resources in all languages.

Agenbroad, J. E. (2006). Romanization is not enough. *Cataloging & Classification Quarterly*, 42(2), 21-34.

Jim Agenbroad is a retired Library of Congress systems analyst. He has long been an advocate for increased non-Latin bibliographic access and was the author of 2 highly debated resolutions within the American Library Association (ALA) pushing for reforms

in this arena. Both of his resolutions were ultimately tabled due to feasibility concerns, but the discussion surrounding them led to the formation of a Task Force on Non-English Access within the Association for Library Collections and Technical Services (ALCTA) – the cataloging and technical services division of the ALA (see Association for Library Collections and Technical Services, 2007). As can be surmised from its title, this article is primarily polemical, though it also gives a good historical overview of romanization in libraries. Agenbroad argues that the reliance on romanization to provide access to non-Latin script materials in our library catalogs and bibliographic databases perpetuates inequality of access to library collections at a time when technological solutions are possible that could expand that access. As he sees the issue, the obstacles to progress in the implementation of non-Latin access points exist more at the level of institutional policy and professional standards than of technological feasibility, and there needs to be concerted effort at the professional level to push for changes. He ends with a couple of recommendations mirroring those he has formally presented to the ALA – to expand the MARC character repertoire and to add rules to the Anglo American Cataloging Rules (AACR2 – or RDA: Resource Description and Access, as it's now becoming) to allow non-Latin access points. (The first of these recommendations has actually come to pass – see Library of Congress, 2007.)

Aliprand, J. (2005). Scripts, languages, and authority control. *Library Resources & Technical Services*, 49(4), 243-9.

Joan Aliprand is a Senior Analyst at the Research Libraries Group (RLG) and was Secretary of the Unicode Consortium from 1991 to 2007. She was also a member of the ALA Task Force on Non-English Access (see Association for Library Collections and Technical Services, 2007). In this article she presents the need for "locale-specific" access points (of which language is one part) and lays out the structural constraints MARC21 places on multilingual and multiscrypt authority records. The technical overview of how multilingual authority files can actually be implemented in MARC is useful in providing background for understanding some of the recent debates and advances in this area. (Association for Library Collections and Technical Services, 2007; Library of Congress, 2007 & 2008).

Association for Library Collections and Technical Services. (2007, March 16.) *Task force on non-English access: Report*. Retrieved March 21, 2008 from <http://www.ala.org/ala/alcts/newslinks/nonenglish/07marchrpt.pdf>

In October 2005, ALCTA – the technical services division of the ALA, created a Task Force on Non-English Access to "examine ALA's past, present, and potential future roles in enabling access to library resources in all languages and scripts and in addressing the needs of users of materials in all languages and scripts through the development of library standards and practices." The task force was created in the wake of a debate over

a resolution authored by Jim Agenbroad pushing for an expansion in the scripts that can be utilized in MARC21 (see Agenbroad 2006, Library of Congress, 2007).

The primary utility of the task force report is probably in its exhaustive survey of the current situation regarding multilingual bibliographic control, at least in the United States. The report includes detailed information from all the major stakeholders in the process – bodies within the ALA, the Library of Congress, OCLC, RLG, and from ILS vendors. Unfortunately, the task force's recommendations are not as robust as the context they provide – providing a framework for future discussion of issues rather than pushing concrete initiatives. There is also little discussion regarding end user input of non-roman scripts, an issue that must be addressed before non-roman script capabilities at the level of the bibliographic record will be of much practical use for most libraries without specialized equipment. That said, this report is certainly an important step, and hopefully will provide direction for future developments. (The report also includes a comprehensive bibliography of resources on Unicode in the library environment – some of which are featured here.)

Kwaśnik, B.H. and Rubin, V.L. (2003). Stretching conceptual structures in classifications across languages and cultures. In N.J. Williamson and C. Beghtol (Eds.), *Knowledge organization and classification in international information retrieval* (pp. 33-47). Binghamton, NY: Hawthorne Information Press.

Barbara H. Kwaśnik is a professor at the Information School at Syracuse University, where she specializes in knowledge organization and research methods. This paper, authored with Syracuse graduate student (and natural language processing researcher) Victoria Rubin, examines cultural and linguistic barriers to translation of classificatory systems. To illustrate their points, the researchers compare kinship classification terminology as codified in Library of Congress Subject Headings (LCSH) and Dewey Decimal System (DDS) schema to kinship classification terminology from different languages and cultures, as determined by ethnographic interviews with native speakers. Their analysis examines differences in scope of terminology, varying lexical and categorical mappings (e.g. empty lexical categories in some languages when compared to others), differences in criteria for distinctions, and differences in cultural usage of kinship terminology. LCSH and the DDS, they argue, by virtue of their cultural origin, encode in their classificatory schema distinctions that apply primarily to English speakers, and as such are less appropriate in other languages and cultures. They conclude with recommendations on how to extend and adapt classification systems in order to better map them to different languages, noting however that while it is possible to increase the utility of translation-classification results, "an ideal translation that is 100

percent culturally and linguistically sensitive is probably not achievable (p. 46)." (For more on problems in mapping general classification systems to specific domains see Mai, 2003.)

Library of Congress. (2007, January) *Character sets and encoding: Part 3: Unicode encoding environment*. Retrieved March 23, 2008 from MARC 21 Specifications for Record Structure, Character Sets, and Exchange Media: <http://www.loc.gov/marc/specifications/speccharucs.html>.

These are the most recent Library of Congress specifications on the use of Unicode in MARC. What is most interesting for our purposes is the following, from the introduction:

To facilitate the movement of records between MARC-8 and Unicode environments, it was recommended for an initial period that the use of Unicode be restricted to a repertoire identical in extent to the MARC-8 repertoire. In 2007, however, such a restriction is no longer appropriate. The full UCS repertoire, as currently defined at the Unicode web site, is valid for encoding MARC 21 records.

This change in the MARC specifications opens up MARC to a greatly increased range of scripts (and therefore languages). Though this marks an important step in the journey towards language-neutral bibliographic control, this is still only part of what is needed. One of the main issues remaining is the lack of library systems that can correctly parse and display these scripts. Hopefully the adoption of the full Unicode character set at this fundamental level should serve to push software developers to enhance their Unicode capabilities.

Library of Congress. (2008, January 10). *White paper: issues related to non-Latin characters in name authority records*. Retrieved March 21, 2008 from <http://www.loc.gov/catdir/cpsd/nonlatin.pdf>.

This paper announces a cooperative agreement among the major authority record exchange partners (the Library of Congress, British Library, National Library of Medicine, and OCLC) to begin adding non-Latin script references in the NACO name authority files. (This is a realization of one of the few concrete recommendations in the Association for Library Collections and Technical Services (2007) task force's report.) Initially, the data will be pre-populated with existing non-Latin headings already present in Worldcat (as many as 500,000 references). Once the references have been added to the authority records they will be analyzed to determine past practices, and this information will be used to develop uniform guidelines and best practices for adding future non-Latin references. The majority of this white paper focuses on providing background on the varieties of existing practice and raising issues that should be considered in the development of the guidelines. This is a major step in the world of multilingual bibliographic control, and it will be interesting to follow the initiative's development.

Mai, J. (2003). The future of general classification. In N.J. Williamson and C. Beghtol (Eds.), *Knowledge organization and classification in international information retrieval* (pp. 3-12). Binghamton, NY: Hawthorne Information Press.

Jens-Erik Mai is a widely published knowledge organization theorist. At the time of this article he was an assistant professor at the University of Washington iSchool. Currently he serves as Vice Dean of the Faculty of Information Studies at the University of Toronto. This article discusses issues involved in establishing mappings between classification systems to provide interoperability. He discusses these issues both in terms of domain specific classification languages (e.g. subject thesauruses) and classification systems from different natural languages. First, Mai makes a distinction between two methods of creating interoperability across collections: the use of an intermediate or switching language, and the use of a general classification scheme in both systems. He goes on to examine each of these domains, focusing on semantic mapping issues and the cultural concerns and theoretical drawbacks (in terms of objectivity and specificity) of different types of classification systems.

While this article only touches briefly on natural languages, the theoretical framework provided is directly applicable to the multilingual classification we are examining, especially when considered in tandem with the literature focusing on mapping conceptual structures in natural language. Mai's discussion of the problems with general classification systems echo Kwaśnik and Rubin's (2003) examination of LCSH and DDC kinship terminology. While Kwaśnik and Rubin advocate the expansion and adaptation of

general classification schemes to better represent different languages, however, Mai argues that any classification system sufficiently general to map to other languages will lack the specificity needed to be useful for access. (Because information retrieval is situated in a context, he says, specific classification languages that reflect that context will be of more utility in retrieving information).

Mai concludes by suggesting that perhaps the best way to provide a degree of interoperability between different classificatory systems is to use general classification systems in conjunction with specific classification systems, where the general classification can be used as a first order organization of knowledge and the specific systems provide further refinement. In the end, though the pragmatic approach of Kwaśnik and Rubin probably makes more sense in terms of making information available most easily in different domains, it is still important to understand the compromises inherent in the use of general classification systems.

MacEwan, A. (2000). Crossing language barriers in Europe: linking LCSH to other subject heading languages. *Cataloging & Classification Quarterly*, 29(1/2), 199-207.

In this article Andrew MacEwan, the Head of Cataloguing at the British Library, describes the MACS (Multilingual Access to Subjects) project, a cooperative effort between the national libraries of Switzerland, France, Germany, and Great Britain. The project aims to establish links between the authority records for the existing controlled vocabularies used for subject access in these libraries. The article reports on a prototype

study with a small subset of the terms (*Sports* and *Theater* terms) intended to establish a methodology for the selection and linking of the headings and to evaluate the utility of these links by comparing the indexing of specific titles in the different systems.

While MacEwan does a good job of explaining the context and rationale behind the MACS program, as well as the methodology of the study, this article doesn't clearly address issues relating to inexact mapping between languages inherent in any translation related enterprise, (as discussed in Kwaśnik and Rubin, 2003; and Mai, 2003). (Note that they are not attempting to map the entire systems – just the authority records. This is equivalent to the approach advocated by Mai in which a general classification is used to bridge specific classification systems.)

Oard, D. W., Diekema, A.R. (1998). Cross-language information retrieval. *Annual Review of Information Science and Technology*, 33, 223-245.

Though not specifically related to multilingual bibliographic control, this article is applicable to providing access to multilingual information – particularly as more and more library information sources are in the form of electronic databases. Cross-language information retrieval (CLIR) is the retrieval of information in one language based on search terms in another language (free-text rather than controlled-vocabulary searching). The expansion of multilingual electronic resources on the World Wide Web has made this a topic of great interest to information researchers in the last decade and a half. This article, by two prominent CLIR researchers (Doug Oard, a information technology researcher from the University of Maryland, and Anne Diekema from the School of

Information Studies at Syracuse University) provides a good overview of the history, theory and practice of CLIR in computerized information systems. Oard and Dikema begin by examining the literature on user needs for CLIR and then proceed to explain and examine the stages of the CLIR process. The review ends with a discussion of evaluation techniques and discussion of future research directions. Though slightly dated, the article does a good job of laying out the issues in a way that is accessible to the non computer scientist. (For those readers who are interested in exploring this area in more depth, I would suggest investigating the publications of the Cross Language Evaluation Forum - [http://www.clef-campaign.org/.](http://www.clef-campaign.org/))

Seely, E. (1993). Cataloging non-English materials at Cleveland Public Library: A one hundred twenty-four year history. *Cataloging & Classification Quarterly*, 17(1-2), 257-65.

In this article Edward Seely, the head of Technical Services at the Cleveland Public Library (CPL) outlines the cataloging process used for non-English materials over the 124 year history of the library. This article provides a useful summary of past and current practices from a library with a large non-English collection (over 200,000 volumes in 45 languages in 1991). As with many American libraries, CPL does no cataloging in non-Latin scripts – all non-Latin scripts are transliterated in cataloging. Seely states that, while cataloging in original scripts would increase access for patrons, the staff (especially multilingual staff) and resources to do so are not available. These

types of practical concerns will need to be addressed in any systematic effort to increase multilingual access to library collections.

Spalding, C.S. (1977). Romanization reexamined. *Library Resources & Technical Services*, 21(1), 3-12.

C. Sumner Spalding was a 35 year employee of the Library of Congress, holding the position of Assistant Director (Cataloging) at the time of his retirement (2 years prior to this article). This article was occasioned by the preparation process of the second edition of the AACR. Spalding wanted to use the revision as an opportunity to re-examine the use of romanization in our catalogs, arguing that such a practice restricted access to materials solely to retain the concept of the "universal catalog," and that abandoning romanization would serve to: (a) increase access for all readers, (b) sidestep the increasingly complex problems caused by conflicting and changing romanization standards, and (c) encourage international cooperation and universal bibliographic control. Spalding proposed a mixed system with separate Author/Title databases by language and a unified subject catalog (though with languages separated out within subjects) as the best approach.

As we know, AACR2 did not end the practice of romanization, and even with the significant progress that has been made on non-Latin cataloging practice since this article we are still a long way from abandoning it completely. Perhaps now that we are on the cusp of another revision (AACR2 into RDA), it is time to reconsider Spalding's argument (as argued by Agenbroad, 2006), especially since the adoption of computerized

cataloging systems rather than card catalogs means that we can do so more easily while still retaining the concept of the "universal catalog" (through multilingual authority control records, etc.).

Tucker, A.M. (1986). Non-Roman and multi-script databases: Basic issues in design and implementation. In C. Boßmeyer and S.W. Massill (Eds.), *Automated systems for access to multilingual and multiscrypt library materials: Problems and solutions* (pp. 34-48). Munich: K.G. Saur.

Alan M. Tucker was the chief designer of the Research Library Group (RLG)'s implementation of Chinese, Japanese, and Korean catalog records in 1983. In this 1986 article he surveys the issues involved with non-Latin bibliographic databases at that time. It is interesting to compare this to the Aliprand article from 2005 to see the extent by which multiscrypt cataloging has been simplified by the advent of Unicode. Much of Tucker's article deals with the issues of handling multiple character sets in a single record, a problem that will fade as more systems transition to Unicode. One thing this article addresses which is curiously lacking from later treatments of this issue is a section dealing with user input in different scripts.

Weih, J. (1998). Interfaces: Three tales of multilingual cataloging. *Technicalities*, 18(10). 6-8.

Jean Weih is a long time cataloging instructor in Canadian library schools and serves as the Canadian Committee on Cataloguing representative on the Joint Steering Committee for Revision of AACR. In this brief article she reviews the history of multilingual cataloging in 3 different countries with significant populations speaking different languages – Canada, South Africa, and Israel. Canada catalogs books either in French or English, depending on their source language (with all other languages cataloged in English). South Africa (with 11 official languages) simply catalogs everything in English, while noting languages in the records. Finally, Israel maintains separate catalogs in Hebrew, Arabic, Cyrillic, and Roman scripts, while subject access is either through a classified catalog, or increasingly through adoption of English LCHS headings. While Weih looks to Unicode integration as a possible way to integrate multiple catalogs, she notes that "until the day arrives when computers can routinely collate bibliographic records and different alphabets, it appears that English has become the collocation device for many libraries."