The Path toward Global Interoperability in Cataloging

Ilana Tolkoff

Libraries began in complete isolation with no uniformity of standards and have grown over time to be ever more interoperable. This paper examines the current steps toward the goal of universal interoperability. These projects aim to reconcile linguistic and organizational obstacles, with a particular focus on subject headings, name authorities, and titles.

In classical and medieval times, library catalogs were completely isolated from each other and idiosyncratic. Since then, there has been a trend to move toward greater interoperability. We have not yet attained this international standardization in cataloging, and there are currently many challenges that stand in the way of this goal. This paper will examine the teleological evolution of cataloging and analyze the obstacles that stand in the way of complete interoperability, how they may be overcome, and which may remain. This paper will not provide a comprehensive list of all issues pertaining to interoperability; rather, it will attempt to shed light on those issues most salient to the discussion.

Unlike the libraries we are familiar with today, medieval libraries worked in near total isolation. Most were maintained by monks in monasteries, and any regulations in cataloging practice were established by each religious order. One reason for their lack of regulations was that their collections were small by our standards; a monastic library had at most a few hundred volumes (a couple thousand in some very rare cases). The "armarius," or librarian, kept more of an inventory than an actual catalog, along with the inventories of all other valuable possessions of the monastery. There were no standard rules for this inventory-keeping, although the armarius usually wrote down the author and title, or incipit if there was no author or title. Some of these inventories also contained bibliographic descriptions, which most often described the physical book rather than its contents. The inventories were usually taken according to the shelf organization, which was occasionally based on subject, like most libraries are today. These trends in medieval cataloging varied widely from library to library, and their inventories were entirely different from our modern OPACs. The inventory did not provide users access to the materials. Instead, the user consulted the armarius, who usually knew the collection by heart. This was a reasonable request given the small size of the collections.[1]

Ilana Tolkoff (ilana.tolkoff@gmail.com) holds a BA in music and Italian from Vassar College, an MA in musicology from Brandeis University, and an MLS from the University at Buffalo. She is currently seeking employment as a music librarian.

This type of nonstandardized cataloging remained relatively unchanged until the nineteenth century, when Charles C. Jewett introduced the idea of a union catalog. Jewett also proposed having stereotype plates for each bibliographic record, rather than a book catalog, because this could reduce costs, create uniformity, and organize records alphabetically. This was the precursor to the twentieth-century card catalog. While many of Jewett's ideas were not actually practiced during his lifetime, they laid the foundation for later cataloging practices.[2]

The twentieth century brought a great revolution in cataloging standards, particularly in the United States. In 1914, the Library of Congress Subject Headings (LCSH) were first published and introduced a controlled vocabulary to American cataloging. The 1960s saw a wide array of advancements in standardization. The Library of Congress (LC) developed MARC, which became a national standard in 1973. It also was the time of the creation of Anglo-American Cataloguing Rules (AACR), the Paris Principles, and International Standard Bibliographic Description (ISBD). While many of these standardization projects were uniquely American or British phenomena, they quickly spread to other parts of the world, often in translated versions.[3]

While the technology did not yet exist in the 1970s to provide widespread local online catalogs, technology did allow for union catalogs containing the records of many libraries in a single database. These union catalogs included the Research Libraries Information Network (RLIN), the OCLC Online Computer Library Center (OCLC), and the Western Library Network (WLN). In the 1980s the local online public access catalog (OPAC) emerged, and in the 1990s OPACs migrated to the Web (WebPACs).[4] Currently, most libraries have OPACs and are members of OCLC, the largest union catalog, used by more than 71,000 libraries in 112 countries and territories.[5]

Now that most of the world's libraries are on OCLC, librarians face the challenge and inconvenience of discrepancies in cataloging practice due to the differing standards of diverse countries, languages, and alphabets. The fields of language engineering and linguistics are working on various language translation and analysis tools. Some of these include machine translation; ontology, or the hierarchical organization of concepts; information extraction, which deciphers conceptual information from unorganized information, such as that on the Web; text summarization, in which computers create a short summary from a long piece of text; and speech processing, which is the computer analysis of human speech.[6] While these are all exciting advances in information technology, as of yet they are not intelligent enough to help us establish cataloging interoperability. It will be interesting to see whether language engineering tools will be capable of helping catalogers in the future, but for now they are

best at making sense of unstructured information, such as the Web. The interoperability of library catalogs, which consist of highly structured information, must be tackled through software that innovative librarians of the future will produce.

In an ideal world, OCLC would be smoothly interoperable at a global level. A single thesaurus of subject headings would have translations in every language. There would be just one set of authority files. All manifestations of a single work would be grouped under the same title, translatable to all languages. There would be a single bibliographic record for a single work, rather than multiple bibliographic records in different languages for the same work. This single bibliographic record could be translatable into any language, so that when searching in WorldCat, one could change the settings to any language to retrieve records that would display in that chosen language. When catalogers contribute to OCLC, they would create the records in their respective languages, and once in the database the records would be translatable to any other language. Because records would be so fluidly translatable, an OPAC could be searched in any language. For example, the default settings for the University at Buffalo's OPAC could be English, but patrons could change those settings to accommodate the great variety of international students doing research. This vision is utopian to say the least, and it is doubtful that we will ever reach this point. But it is valuable to establish an ideal scenario to aim our innovation in the right direction.

One major obstacle in the way of global interoperability is the existence of different alphabets and the inherently imperfect nature of transliteration. There are essentially two types of transliteration schemes: those based on phonetic structure and those based on morphemic structure. The danger of phonetic transliteration, which mimics pronunciation, is that semantics often get lost. It fails to differentiate between homographs (words that are spelled and pronounced the same way but have different meanings). Complications also arise when there are differences between careful and casual styles of speech. Park asserts, "When catalogers transcribe words according to pronunciation, they can create inconsistent and arbitrary records."[7] Morphemic transliteration, on the other hand, is based on the meanings of morphemes, and sometimes ends up being very different from the pronunciation in the source language. One advantage to this, however, is that it requires fewer diacritics than phonetic transliteration. Park, whose primary focus is on Korean–Roman transliteration, argues that the McCune Reischauer phonetic transliteration that libraries use loses too much of the original meaning. In other alphabets, however, phonetic transliteration may be more beneficial, as in the LC's recent switch to Pinyin transliteration in Chinese. The LC found Pinyin to be more easily searchable than Wade-Giles or monosyllabic Pinyin, which are both morphemic. However, another problem with transliteration that neither phonetic nor morphemic schemes can solve is word segmentation—how a transliterated word is divided. This becomes problematic when there are no contextual clues, such as in a bibliographic record.[8]

Other obstacles that stand in the way of interoperability are the diverse systems of subject headings, authority headings, and titles found internationally. Resource Description and Access (RDA) will not deal with subject headings because it is such a hefty task, so it is unlikely that subject headings will become globally interoperable in the near future.[9] Fortunately, twenty-four national libraries of English speaking countries use LCSH, and twelve non-English-speaking countries use a translated or modified version of LCSH. This still leaves many more countries that use their own systems of subject headings, which ultimately need to be made interoperable. Even within a single language, subject headings can be complicated and inconsistent because they can be expressed as a single noun, compound noun, noun phrase, or inverted phrase; the problem becomes even greater when trying to translate these to other languages. Bennett, Lavoie, and O'Neill note that catalogers often assign different subject headings (and classifications) to different manifestations of the same work.[10] That is, the record for the novel *Gone with the Wind* might have different subject headings than the record for the movie. This problem could potentially be resolved by the Functional Requirements for Bibliographic Records (FRBR), which will be discussed below.

Translation is a difficult task, particularly in the context of strict cataloging rules. It is especially complicated to translate among unrelated languages, where one might be syntactic and the other inflectional. This means that there are discrepancies in the use of prepositions, conjunctions, articles, and inflections. The ability to add or remove terms in translation creates endless variations. A single concept can be expressed in a morpheme, a word, a phrase, or a clause, depending on the language. There also are cultural differences that are reflected in different languages. Park gives the example of how Anglo-American culture often names buildings and brand names after people, reflecting our culture's values of individualism, while in Korea this phenomenon does not exist at all. On the other hand, Korean's use of formal and informal inflections reflects their collectivist hierarchical culture. Another concept that does not cross cultural lines is the Korean *pumasi* system in which family and friends help someone in a time of need with the understanding that the favor will be returned when they need it. This cannot be translated into a single English word, phrase, or subject heading. One way of resolving ambiguity in translations is through modifiers or scope notes, but this is only a partial solution.[11]

Because translation and transliteration are so difficult,

as well as labor-intensive, the current trend is to link already existing systems. Multilingual Access to Subjects (MACS) is one such linking project that aims to link subject headings in English, French, and German. It is a joint project under the Conference of European National Librarians among the Swiss National Library, the Bibliothèque nationale de France (BnF), the British Library (BL), and Die Deutsche Bibliothek (DDB). It aims to link the English LCSH, the French Répertoire d'autorité matière encyclopédique et alphabétique unifié (RAMEAU), and the German Schlagwortnormdatei/Regeln für den Schlagwortkatalog (SWD/RSWK). This requires manually analyzing and matching the concepts in each heading. If there is no conceptual equivalent, then it simply stands alone. MACS can link between headings and strings or even create new headings for linking purposes. This is not as fruitful as it sounds, however, as there are fewer correspondences than one might expect. The MACS team experimented with finding correspondences by choosing two topics: sports, which was expected to have a particularly high number of correspondences, and theater, which was expected to have a particularly low number of correspondences. Of the 278 sports headings, 86 percent matched in all three languages, 8 percent matched in two, and 6 percent was unmatched. Of the 261 theater headings, 60 percent matched in three languages, 18 percent matched in two, and 22 percent was unmatched.[12] Even in the most cross-cultural subject of sports, 14 percent of terms did not correspond fully, making one wonder whether linking will work well enough to prevail.

A similar project—the Virtual International Authority File (VIAF)—is being undertaken for authority headings, a joint project of the LC, the BnF, and DDB, and now including several other national libraries. VIAF aims to link (not consolidate) existing authority files, and its beta version (available at http://viaf.org) allows one to search by name, preferred name, or title. OCLC's software mines these authority files and the titles associated with them for language, LC control number, LC classification, usage, title, publisher, place of publication, date of publication, material type, and authors. It then derives a new enhanced authority record, which facilitates mapping among authority records in all of VIAF's languages. These derived authority records are stored on OAI servers, where they are maintained and can be accessed by users. Users can search VIAF by a single national library or broaden their possibilities by searching all participating national libraries. As of 2006, between the LC's and DDB's authority files, there were 558,618 matches, including 70,797 complex matches (one-to-many), and 487,821 unique matches (one-to-one) out of 4,187,973 LC names and 2,659,276 DDB names. Ultimately, VIAF could be used for still more languages, including non-Roman alphabets.[13] Recently the National Library of Israel has joined, and VIAF can link to the Hebrew alphabet.

A similar project to VIAF that also aimed to link authority files was Linking and Exploring Authority Files (LEAF), which was under the auspices of the Information Society Technologies Programme of the Fifth Framework of the European Commission. The three-year project began in 2001 with dozens of libraries and organizations (many of which are national libraries), representing eight languages. Its website describes the project as follows:

> Information which is retrieved as a result of a query will be stored in a pan-European "Central Name Authority File." This file will grow with each query and at the same time will reflect what data records are relevant to the LEAF users. Libraries and archives wanting to improve authority information will thus be able to prioritise their editing work. Registered users will be able to post annotations to particular data records in the LEAF system, to search for annotations, and to download records in various formats.[14]

Park identifies two main problems with linking authority files. One is that name authorities still contain some language-specific features. The other is that disambiguation can vary among name authority systems (e.g., birth/death dates, corporate qualifiers, and profession/activity). These are the challenges that projects like LEAF and VIAF must overcome.

While the linking of subject headings and name authorities is still experimental and imperfect, the FRBR model for linking titles is much more promising and will be incorporated in the soon-to-be-released RDA. According to Bennett, Lavoie, and O'Neill, there are three important benefits to FRBR: (1) it allows for different views of a bibliographic database, (2) it creates a hierarchy of bibliographic entities in the catalog such that all versions of the same work fall into a single collapsible entry point, (3) and the confluence of the first two benefits makes the catalog more efficient. In the FRBR model, the bibliographic record consists of four entities: (1) the work, (2) the expression, (3) the manifestation, and (4) the item. All manifestations of a single work are grouped together, allowing for a more economical use of information because the title needs to be entered only once.[15] That is, a "title authority file" will exist much like a name authority file. This means that all editions in all languages and in all formats would be grouped under the same title. For example, the *Lord of the Rings* title would include all novels, films, translations, and editions in one grouping. This would reduce the number of bibliographic records, and as Danskin notes, "The idea of creating more records at a time when publishing output threatens to outstrip the cataloguing capacity of national bibliographic agencies is alarming."[16]

The FRBR model is particularly beneficial for complex canonical works like the Bible. There are a small number of complex canonical works, but they take up a

disproportionate number of holdings in OCLC.[17] Because this only applies to a small number of works, it would not be difficult to implement, and there would be a disproportionate benefit in the long run. There is some uncertainty, however, in what constitutes a complex work and whether certain items should be grouped under the same title.[18] For instance, should Prokofiev's *Romeo and Juliet* be grouped with Shakespeare's? The advantage of the FRBR model for titles over subject headings or name authorities is that no such thing as a title authority file exists (as conceptualized by FRBR). We would be able to start from scratch, creating such title authority files at the international level. Subject headings and name authorities, on the other hand, already exist in many different forms and languages so that cross-linking projects like VIAF might be our only option.

It is encouraging to see the strides being made to make subject headings, name authority headings, and titles globally interoperable, but what about other access points within a record's bibliographic description? These are usually in only one language, or two if cataloged in a bilingual country. Should these elements (format, contents, and so on) be cross-linked as well, and is this even possible? What should reasonably be considered an access point? Most people search by subject, author, or title, so perhaps it is not worth making other types of access points interoperable for the few occasions when they are useful. Yet if 100 percent universal interoperability is our ultimate utopian goal, perhaps we should not settle for anything less than true international access to all fields in a record.

Because translation and transliteration are such complex undertakings, linking of extant files is the future of the field. There are advantages and disadvantages to this. On the one hand, linking these files is certainly better than having them exist only for their own countries. They are easily executed projects that would not require a total overhaul of the way things currently stand. The disadvantages are not to be ignored, however. The fact that files do not correspond perfectly from language to language means that many files will remain in isolation in the national library that created them. Another problem is that cross-linking is potentially more confusing to the user; the search results on http://www.viaf.org are not always simple and straightforward. If cross-linking is where we are headed, then we need to focus on a more user-friendly interface. If the ultimate goal of interoperability is simplification, then we need to actually simplify the way query results are organized rather than make them more confusing.

Very soon RDA will be released and will bring us to a new level of interoperability. AACR2 arrived in 1978, and though it has been revised several times, it is in many ways outdated and mainly applies to books. RDA will bring something completely new to the table. It will be flexible enough to be used in other metadata schemes besides MARC, and it can even be used by different industries such as publishers, museums, and archives.[19] Its incorporation of the FRBR model is exciting as well. Still, there are some practical problems in implementing RDA and FRBR, one of which is that reeducating librarians about the new rules will be costly and take time. Also, FRBR in its ideal form would require a major overhaul of the way OCLC and integrated library systems currently operate, so it will be interesting to see to what extent RDA will actually incorporate FRBR and how it will be practically implemented. Danskin asks, "Will the benefits of international co-operation outweigh the costs of effecting changes? Is the USA prepared to change its own practices, if necessary, to conform to European or wider IFLA standards?"[20] It seems that the United States is in fact ready and willing to adopt FRBR, but to what extent is yet to be determined.

What I have discussed in this paper are some of the more prominent international standardization projects, although there are countless others, such as EuroWordNet, the Open Language Archives Community (OLAC), and International Cataloguing Code (ICC), to name but a few.[21] In general, the current major projects consist of linking subject headings, name authority files, and titles in multiple languages. Linking may not have the best correspondence rates, we have still not begun to tackle the cross-linking of other bibliographic elements, and at this point search results may be more confusing than helpful. But the existence of these linking projects means we are at least headed in the right direction. The emergent universality of OCLC was our most recent step toward interoperability, and it looks as if cross-linking is our next step. Only time will tell what steps will follow.

References

1. Lawrence S. Guthrie II, "An Overview of Medieval Library Cataloging," *Cataloging & Classification Quarterly* 15, no. 3 (1992): 93–100.

2. Lois Mai Chan and Theodora Hodges, *Cataloging and Classification: An Introduction*, 3rd ed. (Lanham, Md.: Scarecrow, 2007): 48.

3. Ibid., 6–8.

4. Ibid., 7–9.

5. OCLC, "About OCLC," http://www.oclc.org/us/en/about/default.htm (accessed Dec. 9, 2009).

6. Jung-Ran Park, "Cross-Lingual Name and Subject Access: Mechanisms and Challenges," *Library Resources & Technical Services* 51, no. 3 (2007): 181.

7. Ibid., 185.

8. Ibid.

Continued on page 39

International and O'Reilly Media, Web 2.0 refers to the Web as being a platform for harnessing the collective power of Internet users interested in creating and sharing ideas and information without mediation from corporate, government, or other hierarchical policy influencers or regulators. Web 3.0 is a much more fluid concept as of this writing. There are individuals who use it to refer to a Semantic Web where information is analyzed or processed by software designed specifically for computers to carry out the currently human-mediated activity of assigning meaning to information on a webpage. There are librarians involved with exploring virtual-world librarianship who refer to the 3D environment as Web 3.0. The important point here is that what Internet users now know as Web 2.0 is in the process of being altered by individuals continually experimenting with and improving upon existing Web applications. Web 3.0 is the undefined future of the participatory Internet.

3. Clay Shirky, "Here Comes Everybody: The Power of Organizing Without Organizations" (presentation videocast, Berkman Center for Internet & Society, Harvard University, Cambridge, Mass., 2008), http://cyber.law.harvard.edu/interactive/events/2008/02/shirky (accessed Oct. 1, 2008).

4. Ibid.

5. Lawerence Lessig, "Early Creative Commons History, My Version," videocast, Aug. 11, 2008, Lessig 2.0, http://lessig.org/blog/2008/08/early_creative_commons_history.html (accessed Aug. 13, 2008).

6. Elaine Peterson, "Beneath the Metadata: Some Philosophical Problems with Folksonomy," *D-Lib Magazine* 12, no. 11 (2006), http://www.dlib.org/dlib/november06/peterson/11peterson.html (accessed Sept. 8, 2008).

7. Clay Shirky, "Ontology is Overrated: Categories, Links, and Tags" online posting, Spring 2005, Clay Shirky's Writings about the Internet, http://www.shirky.com/writings/ontology_overrated.html#mind_reading (accessed Sept. 8, 2008).

8. Gene Smith, *Tagging: People-Powered Metadata for the Social Web* (Berkeley, Calif.: New Riders, 2008): 68.

9. Ibid., 76.

10. Thomas Vander Wal, "Folksonomy," online posting, Feb. 7, 2007, vanderwal.net, http://www.vanderwal.net/folksonomy.html (accessed Aug. 26, 2008).

11. Thomas Vander Wal, "Explaining and Showing Broad and Narrow Folksonomies," online posting, Feb. 21, 2005, Personal InfoCloud, http://www.personalinfocloud.com/2005/02/explaining_and_.html (accessed Aug. 29, 2008).

12. Shirky, "Ontology is Overrated."

13. Ibid.

14. Michael Arrington, "Exclusive: Screen Shots and Feature Overview of Delicious 2.0 Preview," online posting, June 16, 2005, TechCrunch, http://www.techcrunch.com/2007/09/06/exclusive-screen-shots-and-feature-overview-of-delicious-20-preview/(accessed Jan. 6, 2010).

15. Smith, Tagging, 67–93 .

16. Vander Wal, "Explaining and Showing Broad and Narrow Folksonomies."

17. Adam Mathes, "Folksonomies—Cooperative Classification and Communication through Shared Metadata" (graduate paper, University of Illinois Urbana–Champaign, Dec. 2004); Peterson, "Beneath the Metadata"; Shirky, "Ontology is Overrated"; Thomas and Griffin, "Who Will Create the Metadata for the Internet?"

18. Shirky, "Ontology is Overrated."

19. Peterson, "Beneath the Metadata."

20. Cory Doctorow, "Metacrap: Putting the Torch to Seven Straw-Men of the Meta-Utopia," online posting, Aug. 26, 2001, The Well, http://www.well.com/~doctorow/metacrap.htm (accessed Sept. 15, 2008).

21. Marieke Guy and Emma Tonkin, "Folksonomies: Tidying up Tags?" *D-Lib Magazine* 12, no. 1 (2006), http://www.dlib.org/dlib/january06/guy/01guy.html (accessed Sept. 8, 2008).

22. Shirky, "Ontology is Overrated."

Global Interoperability *continued from page 33*

9. Julie Renee Moore, "RDA: New Cataloging Rules, Coming Soon to a Library Near You!" *Library Hi Tech News* 23, no. 9, (2006): 12.

10. Rick Bennett, Brian F. Lavoie, and Edward T. O'Neill, "The Concept of a Work in WorldCat: An Application of FRBR," *Library Collections, Acquisitions, & Technical Services* 27, no. 1, (2003): 56.

11. Park, "Cross-Lingual Name and Subject Access."

12. Ibid.

13. Thomas B. Hickey, "Virtual International Authority File" (Microsoft PowerPoint presentation, ALA Annual Conference, New Orleans, June 2006), http://www.oclc.org/research/projects/viaf/ala2006c.ppt (accessed Dec. 9, 2009).

14. LEAF, "LEAF Project Consortium," http://www.crxnet.com/leaf/index.html (accessed Dec. 9, 2009).

15. Bennett, Lavoie, and O'Neill, "The Concept of a Work in WorldCat."

16. Alan Danskin, "Mature Consideration: Developing Bibliographic Standards and Maintaining Values," *New Library World* 105, no. 3/4, (2004): 114.

17. Ibid.

18. Bennett, Lavoie, and O'Neill, "The Concept of a Work in WorldCat."

19. Moore, "RDA."

20. Danskin, "Mature Consideration," 116.

21. Ibid.; Park, "Cross-Lingual Name and Subject Access."