The Integration of Information Languages and Interoperability

Presented by:
Tony Olson
Head of Technical Services
Galter Health Sciences Library
Northwestern University
Chicago, Illinois

Presented at the Program:
"Real World Steps to Interoperability in Libraries",
sponsored by the LITA/ALCTS Authority Control in the Online Environment Interest Group, American Library Association Annual Conference, June 16, 2002, Atlanta, Georgia

Contact Information:
Tony Olson
Galter Health Sciences Library
Northwestern University
303 East Chicago Ave
Chicago, IL 60611 U.S.
(312) 503-8125
email: ajolson@nwu.edu

The other presentations have addressed the problems encountered in the merging of different databases and the efforts to enhance the interoperability of different information systems. The fact that these different databases and information systems also use different information languages adds another significant level of complexity to their integration and interoperability.

Before going on, I need to clarify some of the terminology that I will be using in this talk. I would like to use the terminology proposed by Jacques Maniez in his 1997 paper in Knowledge Organization on "Database Merging and the Compatibility of Indexing Languages"[1], In this paper he divides indexing languages into two types: (1) information languages; and (2) natural languages. Information languages include: classification systems, such as Dewey, LC, etc.; and controlled vocabularies such as thesauri (e.g., Medical Subject Headings (MeSH) from the National Library of Medicine, the Art and Architecture Thesaurus (AAT), etc.) and subject heading lists, such as the Library of Congress Subject Headings (LCSH). Most of this presentation will be devoted to the integration of controlled vocabularies, i..e., thesauri and subject headings lists, although the results of some of the projects that I will be describing could be used to integrate all types of information languages, including classification systems.

By their very nature different controlled vocabularies are incompatible. Controlled vocabularies (even general ones such as LCSH) are developed to address fairly specific needs or fairly specific audiences, and to index databases and systems containing records that have been created using a certain set of standards and guidelines. As pointed out by Lancaster[2], that while controlled vocabularies promote consistency within the systems for which they are designed, they tend to reduce intersystem and database compatibility. We are all familiar with the problems that arise when an attempt is made to merge two or more controlled vocabularies in a database, or to search across multiple databases and systems using different vocabularies. Let us review these major problems (using examples from two of the more commonly used vocabularies, LCSH and MeSH).

1. Conflicts between cross references in one vocabulary and established headings in the other vocabularies. (E.g., the LCSH cross reference term, Pharmacology, Clinical conflicts with the corresponding MeSH established heading.)
2. No references or links between corresponding headings from different vocabularies. (E.g., in LCSH there is no reference from Drug Hypersensitivity (which is the MeSH heading) to the LCSH established heading, Drug Allergy.
3. Differences in syntax in the construction of subject heading strings. (E.g., the LCSH heading, Breast--Cancer, corresponds to the MeSH phrase heading, Breast Neoplasms.)
4. Although a substantial majority of the correspondences between terms in different vocabularies may be one-to-one, there is a significant number of correspondences that are not. (E.g., the pre-coordinated LCSH heading, Art therapy for children, corresponds to the post-coordinated MeSH headings, Art Therapy and Child used together in the same record.)
5. Differences in semantic relationships between vocabularies, which in turn also lead to one-to-many correspondences. (E.g., the LCSH heading, Dental surveys, corresponds to 5 MeSH headings, Dental Health Surveys, and the narrower headings, Dental Plaque Index, DMF Index, Oral Hygiene Index, and Periodontal Index.)
6. Identical headings in different vocabularies can cause the retrieval of duplicate entries. This is especially a problem when multiple vocabularies are used in a single catalog or database.

The number of databases and information systems with their associated discordant information languages has increased significantly over the last half century. More recently the technological capabilities of accessing these databases and systems has also increased, beginning with the development of online library catalogs in the 1980's and culminating with the development of the Internet and the Z39.50 standard for the communication and sharing of information. Consequently the problems, as described above, resulting from the incompatibility of information languages have become acute. Or as Maniez has put it, a user attempting to access an ever widening base of knowledge has been met with increasing noise and silence. Several methods have been proposed to make information languages more compatible, or more accurately to integrate (or harmonize) them. Let me first describe briefly some of these methods and some of the projects that have been undertaken in an effort to integrate various information languages. I will then focus on the project that we have been working on at Northwestern for the past twelve years to integrate LCSH and MeSH.

One of the first successful and probably the most widely known projects is the *Unified Medical Language System* (UMLS) from the National Library of Medicine[3]. Begun in 1986 and under continuous development since then, the System is composed of three Knowledge Sources: the UMLS Metathesaurus that integrates over 60 biomedical vocabularies and classifications and links many different names for the same concepts; the SPECIALIST Lexicon that contains syntactic information for many terms; and the UMLS Semantic Network which contains information about the types and categories to which all Metathesaurus concepts have been assigned and the permissible relationships among these types. The purpose of the UMLS is to aid the development of systems that help users retrieve and integrate electronic biomedical information from a variety of sources and to make it easy for users to link disparate information systems.

Another more recent project is the meta-thesaurus developed by Pat Kuhr at the H.W. Wilson Co.[4] She has mapped the subject headings used in 12 different vocabularies covering the various Wilson indexes into a single meta-thesaurus. A user, searching more than one of the indexes at the same time, will be able to enter a term from one of the vocabularies and retrieve not only records indexed with the entry term, but records indexed with equivalent terms from the other vocabularies as well.

The *Multilingual Access to Subjects* (MACS) project[5] (begun in 1997) is an example of integrating multiple subject languages by providing links between equivalent subject headings. Four national libraries in Europe (the Swiss National Library, the Bibliothèque nationale de France, the British Library and the Deutsche Bibiothek) have begun to link equivalent subject headings from the three different subject heading languages, German (SWD), French (RAMEAU) and English (LCSH), used in these national libraries. When the project is completed, a user searching with a subject heading in his or her preferred language will be able to retrieve records from a variety of catalogs that may use different subject heading languages. It is important to note that the MACS project maintains the autonomy and equality of the subject heading languages it intends to integrate. The mapping data (i.e., the links between equivalent terms) is stored in a separate database, and this database is queried to find equivalent subject headings.

Another method of integration is to use a reference language. In this case terms from various information languages are mapped to a term (or classification number) in a single particular information language (called a reference language). A hypothetical example of how a reference language (in this case DDC) might be designed is shown in Figure 1 below.

RL	**IL1**	**IL2**	**IL3**	**IL4**
DDC: 616.994	LCSH: Cancer	MeSH: Neoplasms	UNESCO: Cancer	DDC: 616.994

Figure 1
Example of a hypothetical reference language

The High Level Thesaurus Project (HILT)[6] was a one year project (carried out in the United Kingdom and completed in 2001) to study the problems of incompatibility among various information languages utilized by various libraries and information centers. One of the Project's conclusions was that mapping different information languages offered the best solution to the incompatibility problem. One of the recommendations from the project was to set up a mapping service that would eventually carry-out a mapping of LCSH, the UNESCO thesaurus, AAT, UDC to a DDC backbone (which would serve essentially as a reference language).

Another project that uses DDC as a reference language is the Renardus Project sponsored by the European Union[7]. In this project local classification schemes that are used in subject gateways, are mapped to DDC. The outcome of this project is intended to be a service that can cross-search and cross-browse distributed subject gateways.

The LCSH/MeSH mapping project at Northwestern University is an example of another approach to the integration of controlled vocabularies. As in the MACS project the autonomy and equality of the controlled vocabularies is maintained. However, instead of creating a separate database that contains the linking data, the data is entered into the authority records of the vocabularies being mapped. The LCSH/MeSH mapping project was begun at Northwestern University in 1990 and has continued there up to the present.

When we first started the project, online library catalogs had just recently come into existence. One of the first issues concerning interoperability was encountered in these online catalogs with the merging of two or more controlled vocabularies, the most common being, LCSH, AAT and MeSH. Users of these online catalogs were confronted with the problems that are caused by the incompatibility of different vocabularies. The original goal of our LCSH/MeSH mapping project was to begin to solve these problems by integrating two of the controlled vocabularies in online catalogs. However, it should be readily apparent that the results of our project can be applied to the merging of multiple databases and information systems. In fact it seems logical that we must first solve the problems of the incompatibility between controlled vocabularies within a single catalog or database, before moving on to the interoperability of multiple databases and information systems. So, while the rest of my presentation will focus on the LCSH/MeSH mapping project, keep in mind that the results can be applied to the interoperability of multiple databases and information systems. Furthermore, many of the problems and issues that we have encountered in this project at Northwestern, and the decisions that we have had to make are shared in one form or another by the other mapping projects.

All of the projects described above whose goal is the integration of two or more information languages can essentially be divided into two major and distinctive components (or objectives).

1. Establish equivalencies between terms in the information languages that are to be integrated, and to record these equivalencies (e.g., in authority records, databases, or meta-thesauri). More commonly this activity is referred to as linking or mapping the information languages involved. Henceforth in this talk I will refer to all of these projects as mapping projects.
2. Develop software or enhance existing software, so that the mapping data can be used in catalogs, databases and other information systems that are to be merged.

The first component of the LCSH/MeSH mapping project at Northwestern has been completed. We have developed a combination of computer-assisted techniques and human editorial review, which are used to determine if an LCSH heading corresponds to a MeSH heading, or vice versa. We have mapped corresponding headings by adding 7XX linking entry fields to MARC 21 authority records. These linking entry fields contain headings that correspond to the established headings in 1XX fields. At present as a result of the mapping project, MeSH headings that correspond to LCSH headings have been entered into 750 and 788 linking entry fields in over 11,000 LCSH authority records, and similarly, LCSH headings that correspond to MeSH headings have been entered into 750 and 788 fields of over 9,700 MeSH authority records. These enhanced authority records, with the mapping data are now in the Northwestern online catalog. Also as LCSH and MeSH headings have been added, changed or deleted, we have continued to update the mapping data. A detailed description of this part of project entitled "Mapping the LCSH and MeSH Systems" can be found in the March 1997 issue of *Information Technology and Libraries*.[8] Some examples of LCSH and MeSH authority records containing the mapped headings are shown in Figures 2-4 below.

008/11 a *[code indicating LCSH]*
150: : $a Drug allergy
750: 2: $a Drug Hypersensitivity

008/11 c *[code indicating MeSH]*
150: : $a Drug Hypersensitivity
750: 0: $a Drug allergy

Figure 2
Example of a one-to-one correspondence between an LCSH and MeSH heading, with the corresponding headings entered into 750 linking entry fields.

008/11 a *[code indicating LCSH]*
150: : $a Breast $x Cancer
750: 2: $a Breast Neoplasms

008/11 c *[code indicating MeSH]*
150: : $a Breast Neoplasms
750: 0: $a Breast $x Cancer
750: 0: $a Breast $x Tumors

008/11 a *[code indicating LCSH]*
150: : $a Breast $x Tumors
750: 2: $a Breast Neoplasms

Figure 3
Example of a one-to-two correspondence in which two LCSH headings correspond to a single MeSH heading. This example also shows the mapping of main heading/subheading strings to a main heading.

008/11 a *[code indicating LCSH]*
150: : $a Art therapy for children
750: 2: $8 1 $w b $a Art Therapy
750: 2: $8 1 $w b $a Child
788: 2: $i Search also under the following headings used together in the same record:
    $a Art Therapy $i and $a Child

Figure 4
In this example a pre-coordinated LCSH heading corresponds to two post-coordinated MeSH headings. This relationship is expressed in the 788 complex linking entry field. Note that the LCSH heading is not entered into a 750 field in either of the MeSH authority records, because it does not correspond to these headings individually.

Before moving on to a discussion of the other component of the project (i.e., utilizing the mapping data in online catalogs), I would like to discuss some major decisions that we made as we carried out the project. I think that these might be instructive and illuminate common problems shared by all of the mapping projects that I have mentioned.

After reviewing descriptions of the various mapping projects that I briefly summarized earlier, and considering the problems that we had to solve at Northwestern in the LCSH/MeSH mapping project, it becomes readily apparent that one of the more difficult problems encountered in all of the mapping projects is the one-to-multiple correspondences between headings in different controlled vocabularies. I have already given you several examples of this kind of correspondence. One manifestation of this problem is shown in Figure 4 above in which a single pre-coordinated heading in one vocabulary corresponds to two post-coordinated headings from the other vocabulary.

Another aspect of this problem, which arises from the differing semantic relationships in different vocabularies, is the example mentioned earlier in which one LCSH heading corresponds to five MeSH headings. We could have created five mappings for this situation, but we did not. Instead it quickly became apparent that we could use the syndetic structures of both vocabularies to express these relationships. By creating one mapping, in this case LCSH Dental Surveys to MeSH Dental Health Surveys, a user searching for the LCSH heading, Dental surveys, can also be directed to the corresponding MeSH heading, Dental Health Surveys, and can then be further directed to the narrower MeSH headings. Or conversely, a user searching for the MeSH heading, Periodontal Index, will also be directed to the broader MeSH heading, Dental Health Surveys, and furthermore can then be directed to the corresponding LCSH heading, Dental surveys. There are similar examples in which a MeSH heading may correspond to multiple LCSH headings, some of which represent broader or narrower concepts.

However, there are two problems that have to be resolved in order to use the syndetic structure of the two vocabularies as described above.

1. The broader/narrower term relationships are not explicit in the MeSH authority records as distributed by the National Library of Medicine, but are implicit in the category (or tree) numbers in the 072 fields of the MeSH main heading authority records. In order to make these relationships explicit in an online catalog, a series of computer programs have been written by Gary Strawn at Northwestern that compare 072 fields, and based on the category numbers in those fields, 550 fields, containing the broader and narrower terms, are added to the authority records. Examples of enhanced MeSH authority records are shown below.

```
008/11 c [code indicating MeSH]
072:  : $a E5. $x 318. $x 308. $x 250. $x 300
150:  : $a Dental Health Surveys
550:  : $w g $a Health Surveys $5 IEN-HS
550:  : $w h $a Dental Plaque Index $5 IEN-HS
550:  : $w h $a DMF Index $5 IEN-HS
550:  : $w h $a Oral Hygiene Index $5 IEN-HS
550:  : $w h $a Periodontal Index $5 IEN-HS
750: 0: $a Dental surveys


008/11 c [code indicating MeSH]
072:  : $a E5. $x 318. $x 308. $x 250. $x 300. $x 300
150:  : $a Dental Plaque Index
550:  : $w g $a Dental Health Surveys $5 IEN-HS
```

Figure 5
Examples of enhanced MeSH authority records with broader and narrower headings added in 550 fields.

As an added benefit, since these broader and narrower terms have been added to all MeSH main heading authority records (including those not mapped to LCSH headings), our online catalog at Northwestern now displays broader and narrower references between all MeSH headings that have been used in bibliographic records in the catalog.

2. The syndetic structure of LCSH is not complete. There are only narrower term references in LCSH, but no explicit broader term references. However, we hope and expect this problem to be solved in the near future. Ideally the best solution would be for LC to add the appropriate 550 fields (containing narrower terms) to their LCSH authority records. Alternatively, since the programming would not be too difficult, we have thought about doing this at Northwestern ourselves. There are probably other solutions (such as enhancements to library management

systems software), but adding the references to authority records would seem to be the easiest to implement.

Another major problem encountered by the mapping projects is the syntactical differences between subject heading strings in the various vocabularies. E.g., as mentioned above, the MeSH precoordinated phrase Breast Neoplasms corresponds to the LCSH main heading/subheading string Breast--Cancer. All occurrences of this type of correspondence are mapped in the LCSH/MeSH mapping project. Furthermore, in those cases in which there was no existing LCSH authority record for the heading string, we created an authority record in order to record the mapping. On-the-other hand main heading/subheading to main heading/subheading correspondences are not mapped. So, for example, the MeSH heading string Neoplasms--diagnosis is not mapped to the corresponding LCSH string, Cancer--diagnosis. When we began the project in 1990, there were several reasons for this decision.

1. It would have greatly increased the number of mappings to be recorded manually. We believed that eventually LC would create and distribute subheading authority records. When this happened, we would map LCSH and MeSH subheadings. We could then automatically add mainheading/subheading mappings to authority records using batch change programs.
2. Although all valid MeSH mainheading/subheading strings have authority records, there were no authority records for many of the LCSH main heading/subheading strings.
3. One of the original objectives of the project was to use the mapping data to provide links, via explicit see also references, between corresponding headings in online catalogs. At the time that we began the project, we thought that it would be sufficient to map only main headings (or main heading to main heading/subheading strings), and then use the displays of subject heading strings in online catalogs to lead users to the narrower main heading/subheading strings. (Of course this would not work in systems in which the links are not explicitly displayed. In these cases all correspondences would have to be mapped.)

Let us now discuss the other major component of the LCSH/MeSH mapping project. That is to actually integrate these controlled vocabularies in online catalogs, by utilizing the mapping data in conjunction with software modifications and enhancements to the underlying library management systems (LMS). We would like to complete the second component of the LCSH/MeSH mapping project at Northwestern, and then to extend the results to other library catalogs. The enhancements needed to do this are listed below.

- Index 7XX linking entry fields in authority records. This would at least give us the capability of providing *see also* references between corresponding LCSH and MeSH headings in our online catalog. It would also allow us to manipulate the data in these fields more efficiently. It would seem that this enhancement would not be too difficult to implement, since it would only be an extension of the existing functionality that indexes 4XX and 5XX fields in authority records.
- Provide the capability for dealing with duplicate retrieval, because identical LCSH and MeSH headings are used in the same record. This enhancement might be a little more difficult, because it would probably require the re-structuring of the display of the subject heading index in Northwestern's online catalog.
- Resolve conflicts between cross references and established headings. This too might require the re-structuring of the subject heading index display.

Since we are no longer in the business of writing software for library management systems, we are relying on the vendor of Northwestern's library management system to provide the enhancements. However, even though we have been requesting these enhancements, since migrating to our current system four years ago, they have not yet been forthcoming. Consequently we have not been able to complete this component of the project. However, Northwestern's library management system is not alone in these shortcomings. As far as I can determine, there is no other library management system that indexes or does anything else with 7XX linking entry fields in authority records, nor for that matter does anything to integrate different controlled vocabularies in their catalogs. (There is one possible exception, but even in this case, it appears that the linking entry fields are only displayed in the online catalog as part of the authority record display, but that there are no references, or links, between corresponding headings in different vocabularies.)

It should be pointed out that, with the exception of the Unified Medical Language System, none of the other mapping projects have actually completed the second component, i.e., to utilize the mapping data to integrate different information languages, and present the results to our users. In addition to completing this most important goal, it is also important to complete the second component, because until we can present the results of the various mapping projects to our users, we can not determine the true value and usefulness of these projects. It should be fairly obvious that all six of the mapping projects that I have described, require a significant amount of time, effort and expense. Before embarking on additional projects of this nature (such as mapping AAT to LCSH), we need to determine if this time, effort and expense results in a valuable end product that will be used.

Finally I would like to mention several other potential applications that might emanate from the LCSH/MeSH mapping project.

The mapping data could be used by catalogers to assist in the assignment of MeSH (or LCSH) headings. For example, currently at the Health Sciences Library of Northwestern University, copy catalogers attempting to add MeSH headings to bibliographic records containing only LCSH headings, can find corresponding MeSH headings in our enhanced LCSH authority records.

As noted earlier in this presentation the results could also be applied to the merging of multiple databases and information systems, and the integration of their information languages, if these happen to include LCSH and MeSH. Some examples of these applications might be:

- The data could be incorporated into a higher level integrated vocabulary or meta-thesaurus (such as the UMLS).
- The data could be used as an aid in the development of reference languages (such as envisioned in the HILT project) or other tools for the automatic harmonization of information languages.

In order to make the LCSH/MeSH mapping data available for these kinds applications, we plan to extract files of the enhanced LCSH and MeSH authority records with the mapping data from the Northwestern catalog. We will make them available via anonymous FTP on a server at Northwestern University. Libraries and other information centers can download these records and either add them to their own system, or utilize them in any other manner that they wish to. We hope to have these files of enhanced LCSH and MeSH authority records available for distribution sometime later in the summer of 2002.

[1] Maniez, Jacques. Database merging and the compatibility of indexing languages. *Knowledge Organization*, 24(4) (1997) p. 213-224.

[2] Lancaster, F.W. *Vocabulary Control for Information Retrieval*. (2nd ed.) Arlington, Va.: Information Resources Press, 1989. 270 p.

[3] National Library of Medicine. *Unified Medical Language System: Fact Sheet*. [http://www.nih.gov/pubs/factsheets/umls.html] Accessed 05/16/02.

[4] Kuhr, Patricia S. Putting the world back together: mapping multiple vocabularies into a single thesaurus. In: *Subject Retrieval in a Networked Environment: Papers presented at an IFLA Satellite Satellite Meeting*. Sponsored by the IFLA Section on Classification and Indexing & IFLA Section on Information Technology, held at OCLC, Dublin, Ohio, USA, 14-16 August 2001.

[5] Freyre, Elisabeth & Naudi, Max. MACS: subject access across languages and networks. In: *Subject Retrieval in a Networked Environment: Papers presented at an IFLA Satellite Satellite Meeting*. Sponsored by the IFLA Section on Classification and Indexing & IFLA Section on Information Technology, held at OCLC, Dublin, Ohio, USA, 14-16 August 2001. (A complete description and prototype of this project can be found at: [http://www.infolab.kub.nl/prj/macs]

[6] Nicholson, Denise & Neill, Susannah. Interoperability in subject terminologies: the HILT Project. *New review of Information Networking*. 7 (2001) p. 147-157.

[7] Koch, Traugott. Renardus: cross-browsing European subject gateways via a common classification system (DDC). In: *Subject Retrieval in a Networked Environment: Papers presented at an IFLA Satellite Satellite Meeting*. Sponsored by the IFLA Section on Classification and Indexing & IFLA Section on Information Technology, held at OCLC, Dublin, Ohio, USA, 14-16 August 2001.

[8] Olson, Tony & Strawn, Gary. Mapping the LCSH and MeSH systems. *Information Technology and Libraries*. 16(1) (March 1997) p. 5-19.