Date : 17/06/2008

Subject Headings for the 21st Century: The lcsh-es.org Bilingual Database

Michael Kreyche
Systems Librarian, Kent State University
Kent, Ohio, USA

Meeting: **129. Classification and Indexing**
Simultaneous Interpretation: English, Arabic, Chinese, French, German, Russian and Spanish

WORLD LIBRARY AND INFORMATION CONGRESS: 74TH IFLA GENERAL CONFERENCE AND COUNCIL
10-14 August 2008, Québec, Canada
http://www.ifla.org/IV/ifla74/index.htm

Abstract

Spanish is one of the most widely spoken languages in the world and a review of the lists of subject headings in this language reveals numerous efforts over a period of time, usually involving some form of collaboration, but largely isolated from each other. Technological developments suggest that a greater degree of cooperation is now possible and would be beneficial to the international library community if other barriers can be surmounted. The lcsh-es.org project demonstrates this concept in a practical way and suggest a new model for international cooperation in authority control. The site may be accessed at http://lcsh-es.org.

Spanish Language and Subject Headings

The language commonly known as Spanish (Español) is one of the most widely spoken and geographically distributed. Estimates for the number of people who speak it vary; the *Encyclopedia Britannica* puts it at 358 million[1] and *Ethnologue*[2] lists 27 countries where it is spoken by 322 million people. It is often referred to as Castilian (Castellano), reflecting its origin as one of several regional dialects and related languages from the Iberian peninsula.

The largest number of Spanish speakers today live in the Americas, extending through most of South America, parts of the Caribbean, Central America, Mexico, and into North America. The migration of Spanish speakers into the United States during the last twenty years is especially dramatic; in large national chain stores it is now the norm to see signs in both English and Spanish. Not surprisingly, serving Spanish speakers is becoming a high priority for libraries in the United States. Many of them offer Spanish pages on their web sites and a Spanish user interface for their catalogs, but it is not so common for the catalogs to contain records in Spanish or records with Spanish subject headings. Unfortunately, there is no single source of headings that is comprehensive, technologically advanced, or well-suited to quickly find equivalents to Library of Congress headings.

The earliest source of published subject headings in Spain, as found in OCLC, is a list for public libraries first published in 1986.[3] The fourth edition of a printed list of headings from a network of research libraries was published by the Consejo Superior de Investigaciones Científicas (CSIC) in 1996[4] and the fifth edition was published in 2000 on CD-ROM. Apparently the authority file of the Biblioteca Nacional de España (BNE) was first published in 1996,[5] also in CD-ROM format[6]. In recent years both the BNE and the CSIC have made their authority files publicly available online.

Excellent information on Spanish subject heading lists developed in the Americas is readily available in a recently published bilingual volume containing papers from a program at the annual conference of the American Library Association (ALA) in 2004, as well as two additional invited papers.[7] Four of the articles discuss one or more of the three most influential lists based on the Library of Congress Subject Headings (LCSH).

One of the lists is the Lista de *Encabezamientos de Materias para Bibliotecas* (LEMB), originally sponsored by the Panamerican Union (now known as the Organization of American States) and first published in 1967. It is maintained today by the Biblioteca Luis Ángel Arango of the Banco de la República in Colombia and is available commercially in electronic format.

Also first published in 1967 was the *Lista de Encabezamientos de Materia* (LEM), commonly referred to as the "Escamilla list" after its compiler, Gloria Escamilla of the Biblioteca Nacional de México. Although the list is still maintained in the library, the only published update thus far was a second edition that appeared in 1978. A proposal for developing a new edition was described at the 70th IFLA General Conference in Buenos Aires by Filiberto Felipe Martínez Arellano[8] from the Universidad Nacional Autónoma de México.

The third list is *Bilindex*, first published as a publicly funded project in 1986. It survives as a commercial venture under the imprint of Floricanto Press. Preparation of the original edition relied on the LEMB and the LEM as well as many other sources, and it should also be mentioned that an early, pre-publication version of the LEM was provided to the original editors of the LEMB. This openness to sharing has suffered in subsequent years, partly due to the commercialization of the LEMB and *Bilindex* and the growing interest in protecting intellectual property rights. Another factor may have been the closed nature of the automated systems adopted in the 1980s and 1990s for maintaining authority data.

There are currently two noteworthy collaborative projects in Mexico, one led by the Colegio de México and mentioned by Martínez Arellano in his *SALSA* presentation. It is fairly well-known and has been written about elsewhere.[9] The other is relatively new and was described in detail by the coordinator of the project, Julia Margarita Martínez Saldaña of the Universidad Autónoma de San Luís Potosí at the same program. The major focus of the collaboration is participation in the Name Authority and Subject Authority Cooperative Programs (NACO and SACO[10]) of the Library of Congress rather than Spanish subject headings. Development of an interface for exchanging authority records between three different systems has overcome some of the limitations of a closed system.

The 2004 IFLA meeting in Buenos Aires had the effect of stimulating Latin American interest in *Resource Description and Access*, the cataloging rules scheduled for adoption in 2009, and also inspired a series of annual meetings for catalogers. Besides fostering discussion on the new rules, they have covered other cataloging topics, including authority control. The first *Encuentro Internacional de Catalogadores*, held in Lima, Peru in 2005, included several presentations, a panel discussion, and a workshop on authority control.[11] The papers are available in a printed volume.[12] The second *Encuentro* occurred in Mexico City in 2006, where the lcsh-es.org project was first announced. The proceedings of that conference have also been published.[13] The third *Encuentro* was held in Buenos Aires in November of 2007 and featured eight papers on the theme of authority control for names and subjects, all of which have been published on the conference web site.[14]

In the United States, the San Francisco Public Library has long been noted for the care it has taken to include Spanish subject headings in its catalog, which has served as a reference for many catalogers. More recently the Queens Borough Public Library in New York has undertaken a project to create its own list of Spanish subject headings, much of it based on the San Francisco work.

Thanks to a partnership with these two libraries and the good will of the BNE and the CSIC, the lcsh-es.org project is building a bilingual database of subject headings in the hopes of laying a foundation for broader international cooperation on Spanish subject authority work. Of all the obstacles to such a goal—technical, linguistic, geographic, political, and economic—the easiest to overcome seems to be the technical, so that has been the primary focus. The hope is that modern tools designed for the purpose will help overcome the other obstacles.

Technological Environment

Machine readable—and therefore shareable—authority data has been in existence for some twenty-five years. For the most part, however, authority files exist within the context of the catalog of a single library or a group of libraries, making it difficult to share work outside of that context. Likewise, library catalogs have traditionally been closed, proprietary systems designed to perform very specific tasks within that system. That sort of environment is rapidly becoming obsolete.

Over the last decade the open source community has developed an amazing assortment of high quality, general purpose software tools that operate on different hardware platforms. At the same time, technologists have worked hard to develop open standards for data formats and communications protocols. The result is a proliferation of creative experiments often grouped loosely under the term “Web 2.0.”

At present the library community is directing a great deal of energy towards the “Next Generation Catalog”, inspired by the success of the Koha[15] and Evergreen[16] systems and stimulated by the 2007 CODE4LIB conference which featured a pre-conference workshop on the Lucene and SOLR software tools[17] and a program by Casey Durfee entitled, “Open-Source Endeca in 250 Lines or Less”[18] (Endeca being a commercial product used to build some of the new-style catalogs).

While these systems are designed with the library patron in mind, there are opportunities to employ these technologies to serve the special needs of librarians, as in the case with lcsh-es.org. Another example is the Library of Congress Subject Headings database built by Bernhard Eversberg of the Technische Universität Carolo-Wilhelmina zu Braunschweig.[19] It is somewhat ironic that structured vocabularies have caught the attention of researchers interested in building the semantic web while some librarians are questioning their value in relation to the cost of maintaining them. The fact that this objection has some validity is all the more reason to rethink authority work and look to technology for more efficient ways to perform it.

The lcsh-es.org Project

The lcsh-es.org database is based on the premise that conditions are ripe for a new era of international cooperation in authority work, and a new sort of tool is necessary to bring this about, especially for Spanish subject headings. Over a period of time, the characteristics of such a tool became evident. Three of them emerged as immediate goals in the initial phase of development, from September 2005 to September 2007:

Comprehensive: Among the various languages with subject headings systems based on the Library of Congress Subject Headings, Spanish is probably unusual because it has such a wide geographical distribution, many national and regional variations, and thus not one or two but five well established subject systems (BNE, CSIC, LEM, LEMB, and Bilindex) with some new variations or amalgams emerging. Leaving aside for the moment whether this fragmentation is desirable or not, it is clear that consulting multiple sources is a necessity in Spanish subject authority work. So bringing together as many headings from multiple sources is an important aim.

Efficient and easy to use: Comprehensiveness implies efficiency, but making one's job easier and faster is important in its own right. One of the lessons of the Internet is that simplicity and ease of use encourage the adoption of web tools in their early stages when they may be deficient in other respects. In lcsh-es.org, searching is kept very simple and results are presented immediately. Both the English and the Spanish terms are indexed, and basic keyword searches are possible.

Freely available on the web: This is closely related to efficiency and ease of use. To be widely accepted, the only requirement should be a browser, and like other services on the web, especially startups, there is an expectation that it will be available without cost.

Based on three measures of success, lcsh-es.org has proven itself in this first phase. One measure is the number of sources incorporated. Initial data came from bibliographic records from the San Francisco Public Library, which has a long-standing reputation for the care it has taken to put Spanish subject headings into its catalog, dating from its participation in the original Bilindex project. Over the years many catalogers from other libraries have consulted its catalog to find Spanish headings. Matching Spanish headings to corresponding English ones produced a dictionary that made it much easier to look up Spanish equivalents to Library of Congress headings. The second source was the Queens Borough Public Library in New York, which was in the process of creating its own dictionary of Spanish headings. Next came headings extracted from the CSIC CD-ROM and others downloaded from the BNE catalog. Lastly, terms scanned from the English-Spanish index of the original Bilindex brought the number of records to over 50,000. Another useful source of data was the Fred 2.0[20] copy of the LCSH file downloaded and made publicly available by Simon Spero (and also used by Eversberg in the project cited above). Although it did not contribute new Spanish terms to the database, it made it possible to validate the LCSH terms found in the other sources.

The second measure of success is a steadily growing number of users. All of the top 40 institutions or networks identified from the web server's logs during the year 2007 accessed at least 100 pages, with the top 12 accessing more than 1000 pages each. These top 40 users accounted for about 35,000 pages, with Mexico and Spain (mostly academic libraries) accounting for about 10,000 each and the United States (mostly public libraries) the remaining 15,000. Even though the database is still far from being comprehensive, it clearly has achieved a critical mass that makes it useful to a variety of libraries. The activity in Spain was something of a surprise since virtually all of the identifiable institutions are in Catalonia, where Spanish is not the first language. One possible explanation is that librarians there are using the database to look up LCSH terms and then translating the Spanish equivalents to Catalan. If so, this represents an opportunity to make the database trilingual!

The third measure of success is that the initial work helped obtain funding from the National Endowment for the Humanities (NEH) to support the current phase of development (September 2007-August 2008). Most of the funds are being used to speed up development by hiring a student programmer. The principal aim of this phase is to give the database two additional characteristics:

MARC-based: The data is being converted to MARC21 format and the database structure has been modified accordingly. These changes will make output more immediately useful and will facilitate the incorporation of MARC data from additional sources. Thus far (April 2008) much of the NEH-supported work has concentrated on this aspect.

Collaborative: The database must have an infrastructure for collaboration and interaction and a login system has been developed to support this kind of functionality. Users of the database will be able to contribute to in a variety of ways: identifying errors, suggesting changes, and adding 4xx terms or 670 and 675 fields. It should be possible to create new authority records by exchanging 1xx and 4xx terms in an existing record to reflect regional or national usage. The goal is to foster a community of catalogers that can support each other's authority work. With enough participants, a network like this could become a case of using the techniques of tagging and folksonomies to serve a structured vocabulary.

At the same time, some additional features are making the database easier to use and more functional. For example, "See" (and optionally, "See also") references are now incorporated in the database; and highlighting a heading for copying and pasting is as simple as a mouse click (it is no longer necessary to drag the pointer across the heading first). There are plans to add links to dictionaries and other Internet reference tools and a simple export function for downloading individual MARC records is being tested. A function for sending records directly to a local system in the same way that catalogers can export records from OCLC may also be added.

A number of personalization features are also contemplated. One is to permit the cataloger to designate preferred sources of headings for customized search results, and another is to set character encoding and MARC content designation to match that of the local system for copying and pasting.

In addition to improving the user interface for manual work, batch and web service capabilities have the potential to be useful. For example, a library without Spanish headings might upload a file of bibliographic records and schedule a batch process to search the English headings in the database and add equivalent Spanish terms. A web service designed to perform the same function on a heading-by-heading basis could automate this process for new cataloging records entering a system. This would require custom programming on the local library system and perhaps a library using an open source catalog could be persuaded to develop and contribute this functionality to a wider community. Such a service might also prove useful on some non-library web sites.

Closing Thoughts

Looking ahead, the hope is that informal collaboration fostered by the lcsh-es.org project will develop into formal commitments and agreements to contribute data on an ongoing basis. In order to bring together all the major sources of Spanish subject headings will almost certainly require some substantial initial funding to satisfy the needs of the commercial producers. It’s not unrealistic to think that a partnership of a few major libraries could succeed in attracting such financing. The greater challenge will be to develop a model for permanent economic sustainability.

Even if this proves unattainable, the project may inspire librarians working in other languages to adopt similar techniques, perhaps even the same code base. It may be more practical, in fact, to start developing an LCSH-based vocabulary in a language where none already exists or is not very far along. The system uses the Apache web server, the MySQL database and the PHP script language, all of which can run on a number of operating systems. The original database used the Latin-1 character set, but the current version stores all the records in UTF-8, so potentially it could be adapted to virtually any script and any language.

In closing, some fundamental issues deserve a little reflection. First among them should be the shortcomings of LCSH. Its cultural and linguistic biases have long been recognized and its inconsistent practices are the source of much frustration. A fresh, systematic look at these might yield some good ideas for creating a single internationalized or universal Spanish subject heading system. Perhaps a sort of “bilingual” LCSH could be developed, by dividing it into two copies, with a new international version developing from one, cross referenced to the original. The original could continue to develop in a limited way to as long as it is found useful, and then either become frozen or cease to exist.

Another question that suggests itself is whether the idea of an “authority” file still makes sense. Using a single “authorized” term for a concept certainly made sense when catalogs were printed on cards, in books, or even on microfilm. Now that storage is the cheapest computing resource available, perhaps the 1xx fields of the authority record should be eliminated and its terms tagged as 4xx fields, effectively creating a cluster of synonyms which could be qualified on the basis of geographic usage or realm of knowledge. There is no technical reason not to include multiple terms for the same concept in each appropriate bibliographic record. There is no functional reason either, since a search on any one of the synonyms would retrieve all the records. As far as our library patrons are concerned, the "right" terms are the ones they choose.

In any case, the future of subject heading is likely to be vastly different than the past and the present. When books are routinely published in electronic format, subject analysis as we now know it in libraries may cease to exist. As the quantity of electronic texts increase, we can expect to see sophisticated computer software perform the work with perhaps just a little help from librarians or subject specialists. Full text searching is clearly a very effective technique, but the human tendency to categorize and classify is very powerful, so structured vocabularies are likely continue in some form or other.

[1] "Spanish language." Encyclopædia Britannica. 2008. Encyclopædia Britannica Online. 7 Apr. 2008 http://search.eb.com/eb/article-9068992 (accessed April 9, 2008).

[2] Gordon, Raymond G., Jr. (ed.), 2005. Ethnologue: Languages of the World, Fifteenth edition. Dallas, Tex.: SIL International. Online version: http://www.ethnologue.com/show_language.asp?code=spa (accessed April 9, 2008).

[3] Lista de encabezamientos de materia para las bibliotecas públicas. Madrid: Ministerio de Cultura, Dirección General del Libro y Bibliotecas, 1986.

[4] Unidad de Coordinación de Bibliotecas (C.BIC). Lista de Encabezamientos de materia de la Red de Bibliotecas del CSIC.,4a. ed. Madrid: Consejo Superior de Investigaciones Científicas, 1995.

[5] "Autoridades de la Biblioteca Nacional de España." Correo Bibliotecario. Valladolid: Biblioteca de Castilla y León. Sumario 12, enero-febrero 1997. http://www.bcl.jcyl.es/CORREO/plantilla_neditorial.php?id_neditorial=63&id_seccion=5&RsCorreoNum=12 (accessed April 9, 2008).

[6] Autoridades de la Biblioteca Nacional. Madrid: Chadwyck-Healey España, 1996.

[7] Miller, David and Filiberto Felipe Martínez Arellano, editors. Salsa de tópicos = Subjects in SALSA: Spanish and Latin American subject access. Chicago: Association for Library Collections & Technical Services, American Library Association, 2007. ALCTS papers on library technical services & collections; # 14.

[8] Martínez Arellano, Filiberto Felipe. Desarrollo de una lista de encabezamientos de materias en Español. Buenos Aires: World Library and Information Congress, 70th IFLA General Conference and Council, 22-27 August 2004. http:// www.ifla.org/IV/ifla70/papers/039s_trans-Arellano.pdf (English: http://www.ifla.org/IV/ifla70/papers/039e-Arellano.pdf; French: http://www.ifla.org/IV/ifla70/papers/039f_trans-Arellano.pdf; Russian: http://www.ifla.org/IV/ifla70/papers/039r_trans-Arellano.pdf, (accessed April 9, 2008).

[9] Quijano-Solís, Alvaro, Pilar María Moreno-Jiménex, Reynaldo Figueroa-Servín. "Automated Authority Files of Spanish-Language Subject Headings." The LCSH Century: One Hundred Years with the Library of Congress Subject Headings System. Haworth Press, 2000, p. 209-223.

[10] See also the following paper from IFLA70: Cristán, Ana. The SACO Program in Latin America. Buenos Aires: World Library and Information Congress, 70th IFLA General Conference and Council, 22-27 August 2004. http://www.ifla.org/IV/ifla70/papers/040e-Cristan.pdf; German: http://www.ifla.org/IV/ifla70/papers/040g_trans-Cristan.pdf; Russian: http://www.ifla.org/IV/ifla70/papers/040r_trans-Cristan.pdf; Spanish: http://www.ifla.org/IV/ifla70/papers/040s_trans-Cristan.pdf (accessed April 11, 2008).

[11] Conference program, http://bvirtual.bnp.gob.pe/inscripcion/programacion.htm (accessed April 12, 2008).

[12] Centro Bibliográfico Nacional . Nuevas tendencias en la normalización y sistematización de la información : ponencias y conclusions. Lima: Biblioteca Nacional del Perú, Fondo Editorial, 2006.

[13] Martínez Arellano, Filiberto Felipe and Ariel Alejandro Rodríguez García, compliers. Memoria del Segundo Encuentro Internacional de Catalogación : tendencias en la teoría y práctica de la catalogación bibliográfica, 12 al 14 de septiembre de 2006. México : UNAM, Centro Universitario de

Investigaciones Bibliotecológicas, Instituto de Investigaciones Bibliográficas; Library Outsourcing Service, 2007

[14] III Encuentro de Catalogadores. http://www.bn.gov.ar/ACT_EjesEncuentro.aspx (accessed April 12, 2008).

[15] http://www.koha.org/ (accessed April 12, 2008).

[16] http://open-ils.org/ (accessed April 12, 2008).

[17] http://code4lib.org/node/139 (accessed April 12, 2008).

[18] http://code4lib.org/2007/durfee (accessed April 12, 2008).

[19] http://www.biblio.tu-bs.de/db/lcsh (access ed April 12, 2008).

[20] Spero, Simon. Fred 2.0: Cosmos, Taxis, and the Future of Bibliographic Control. http://www.ibiblio.org/fred2.0/wordpress/ (accessed April 12, 2008).