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

World Library and Information Congress
74th IFLA General Conference and Council
Québec, Canada
August 13, 2008

Michael Kreyche
Systems Librarian
Kent State University
mkreyche@kent.edu

KENT STATE
UNIVERSITY

Today I will be summarizing the main points of the formal paper already available on the IFLA web site and the conference CD-ROM. I'll also provide some additional details and give you a little demonstration of the database. After the conference I will post a link to this Powerpoint presentation on the lcsh-es.org site.

Introduction

- Bilingual database of subject headings
- Available on the web
- Spanish and English
- Library of Congress system (LCSH)
- Multiple sources of headings
- Test some new ideas
- Immediate practical use

KENT STATE
UNIVERSITY

This presentation is about a bilingual database of subject headings that is on the web. There are English headings from the Library of Congress and Spanish headings from multiple sources, all based on the Library of Congress system. The purposes of the database are to test some new ideas for managing subject authorities and to provide a tool that is of immediate practical use to catalogers.

Who will be interested?

Catalogers working with:

- Subject heading systems, especially:
  - in a collaborative environment
  - with systems based on LCSH
  - with multilingual subject headings
  - with Spanish language subject headings

Others: linguists, systems designers

Who might be interested in the database? First of all, any cataloger working with subject headings, especially in a collaborative environment. Because the database is on the Web it is readily available anywhere there is internet access and it is neutral in relation to local cataloging systems. The database includes a copy of LCSH and is especially useful for finding Spanish equivalents to LC headings already present in a bibliographic record, which is the usual case in copy cataloging, whether the record comes directly from the LC catalog or from other sources, such as OCLC. The database may also be of interest to linguists interested in language variations from place to place. I especially hope that designers of next generation library systems will take notice and help define web services that can automate much of the work of integrating multilingual subject headings into catalog records.

Spanish—Español—Castellano

- Spoken by more than 300 million people
- Spoken in 27 countries
  - Europe
  - The Americas
  - a few other places
- Regional and national variations
- Various lists of subject headingss

Spanish is one of the most widely spread languages in the world. From its origin it Spain it spread to the Caribbean and Mesoamerica, then through nearly all of South America and northward into what is now the United States. Modern migrations, especially in the last twenty years, have carried the language to all parts of the United States, both to rural and to urban areas. It is so pervasive that it is normal to see bilingual signs in large national chain stores, even in areas with few Spanish speakers. The presence of Spanish is evident in Canada, too. Last week in Montreal I had no trouble finding a restaurant where I could speak Spanish and I even found some small local newspapers in Spanish.

Languages with Similar Profiles

Language	Speakers	Countries
Spanish	322,300,000	24
English	309,400,000	14
Arabic	323,000,000	24
Russian	145,000,000	14
French	79,600,000	6

From the Wikipedia article, "Ethnologue list of most spoken languages"

http://en.wikipedia.org/wiki/Ethnologue_list_of_most_spoken_languages

KENT STATE
U N I V E R S I T Y

IFLA, August 13, 2008

M. Kreyche - 5

Here are some other languages that have some of the same characteristics as Spanish: spoken by many people in many countries. The numbers of countries where the languages are spoken are open to discussion, but these figures are useful for comparative purposes. You can refer to the Wikipedia article or the Ethnologue web site for more details.

Spanish in U.S. Libraries

- Most have some Spanish language books
- Many have web pages in Spanish
- Some have a Spanish catalog interface
- Few have Spanish subject headings
  - Multiple sources of headings
  - Lack of Spanish language expertise
  - Existing sources not suitable for automation
  - Costly (especially for small libraries)

What is the status of Spanish in U.S. libraries? It's not uncommon for libraries to own at least a few Spanish language books and other materials. Libraries that serve Spanish speaking populations frequently have web pages in Spanish, usually with a focus on library programming or other community resources such as classes in English as a Second Language. Libraries with large numbers of Spanish speaking patrons frequently have a catalog with a Spanish interface. This usually represents a significant one-time investment in time and/or money to configure, but it is of limited use without suitable Spanish language terms in the cataloging records. Unfortunately, Spanish words are frequently restricted to the title and perhaps a content note. Subject headings are usually left to chance; they may or may not appear in catalog records from OCLC or other sources. Furthermore, cataloging has become such an automated process that many records are never examined or receive minimal attention. Even libraries committed to including Spanish subject headings in their records face a difficult challenge; there is no single, comprehensive source of headings and it is time-consuming to consult multiple sources. Furthermore, none of the existing sources are well suited to computerized cataloging systems.

Spanish Headings: Spain

- Biblioteca Nacional de España
  - Authority file online, freely available
  - MARC authority display, but no download
  - LCSH terms in 670 fields, cannot be searched
- Consejo Superior de Investigaciones Científicas
  - Authority file online, freely available
  - MARC authority display, but no download
  - LCSH terms in 7xx fields, keyword search

KENT STATE
UNIVERSITY

Spain has two important sources of subject headings—the Biblioteca Nacional and the Consejo Superior de Investigaciones Científicas (roughly translated as the National Council for Scientific Research). Both libraries maintain MARC authority files online for anyone to access via the web. The records can be displayed with the MARC tagging, but cannot be downloaded as proper MARC files for importing into other systems. Records from the Biblioteca Nacional have Library of Congress terms in 670 fields, but they are not searchable. The authority file of the Consejo has LC terms in 7xx fields; these are searchable by keyword. This file is larger than that of the Biblioteca Nacional and includes many specialized terms. The two systems do not always use the same Spanish term.

Spanish Headings: Colombia

- LEMB Digital (Lista de Encabezamientos de Materia para Bibliotecas)
  - Early print version funded by the Panamerican Union
  - Biblioteca Luís Angel Arango
  - Commercial product
  - CD-ROM and online
  - MARC authority display (upper case)
  - Text download

The LEMB digital comes from Colombia and is considered the most comprehensive source of Spanish subject headings. It seems to be the preferred source in South America. It is based on an early printed list funded by the Panamerican Union (now known as the Organization of American States). The Biblioteca Luís Angel Arango currently maintains the list in the form of a MARC authority file and markets a commercial product in CD-ROM and online formats. The records can be displayed with MARC tagging, but as far as I have seen from demonstrations, the headings appear entirely in upper case letters and it is not possible to download MARC records for importing into other systems.

Spanish Headings: Mexico

- Biblioteca Nacional de México
  - Lista de Encabezamientos de Materia (Escamilla)
  - Print format, outdated (1978)
- CCA (Colegio de México)
  - MARC format
  - Indirect access through catalogs
- PCC (Univ. Autónoma de San Luís Potosí)
  - MARC format
  - Not publicly available

A great deal of authority work has taken place in Mexico, but for the most part it is not readily available. An important printed list from the Biblioteca Nacional is now forty years old. In recent years the Colegio de México and a few other academic libraries formed a consortium for the purposes of sharing authority work. They maintain a MARC authority file, but the only outside access to their headings is through the bibliographic records in the libraries' catalogs. The library system of the Universidad Nacional Autónoma de México, one of the members of the consortium, also has an extensive historical file of Spanish headings with LC equivalents, but it's not in MARC format and is only used internally as a resource for current authority work. Another consortium, formed more recently, lead by the Universidad Autónoma de San Luís Potosí. Its major focus at this time is participation in the Name Authority and Subject Authority Cooperative Programs of the Library of Congress and, as far as I know, none of its Spanish headings are publicly available.

Spanish Headings: United States...

- Bilindex
  - Original version produced with public funding
  - Commercial product of Floricanto Press
  - Text (PDF) format (CD-ROM and online)
- San Francisco Public Library
  - Based on original edition of Bilindex
  - No authority records, headings in catalog
  - Used by many libraries

The best known source of headings in the United States is Bilindex, originally produced in the early 1980s with public funding administered by the California State Librarian. The first printed edition was published in 1986 from a MARC file. At that time the only way to maintain a MARC database was on expensive mainframe computers, so when funding for the project ended, the data was stored on magnetic tape. Eventually the tapes became obsolete and unreadable, so this valuable resource was lost. Today Bilindex continues as a commercial product using Adobe's Portable Document Format (PDF). This is technically an electronic format, but has many of the characteristics of print and would require a huge amount of work to convert to MARC.

The San Francisco Public Library has been maintaining subject headings in its bibliographic records for many years but has no Spanish authority records. Most of its headings are based on the original printed versions of Bilindex and many libraries have relied on the San Francisco catalog for their subject headings.

...Spanish Headings: United States

- Westchester Library System (New York)
  - MARC authority records
  - Headings available indirectly through catalog
- Queens Borough Public Library (New York City)
  - Bilingual file of main headings and subdivisions
  - Special XML file with some MARC coding
  - Custom software to process bibliographic records

KENT STATE
U N I V E R S I T Y

IFLA, August 13, 2008

M. Kreyche - 11

More recently, the Westchester Library System has built a MARC authority file of Spanish headings, mostly derived from San Francisco and Bilindex. So far, the only public access to the records is through the library system's public catalog.

A few years ago the Queens Borough Public Library in New York City began to build a machine readable file of the LC subject headings found in the records for their Spanish language collections. Spanish equivalents were found for each main heading and subdivision and added to the file. These are not authority records, but they do contain MARC tags and subfield codes, and serve somewhat the same purpose. Custom software contracted for by the library uses this file to process bibliographic records with LC headings and automatically generate Spanish headings. The software is open source, available to anyone who wishes to use it.

Salsa de Tópicos

- American Library Association program, 2004
- International
  - United States
  - Mexico
  - Colombia
- Bilingual publication, 2007

ALCTS PAPERS ON LIBRARY TECHNICAL SERVICES & COLLECTIONS #14

SALSA de Tópicos | Subjects in SALSA

SPANISH AND LATIN AMERICAN SUBJECT ACCESS

David Miller and Filiberto Felipe Martinez Arellano, Editors

KENT STATE UNIVERSITY
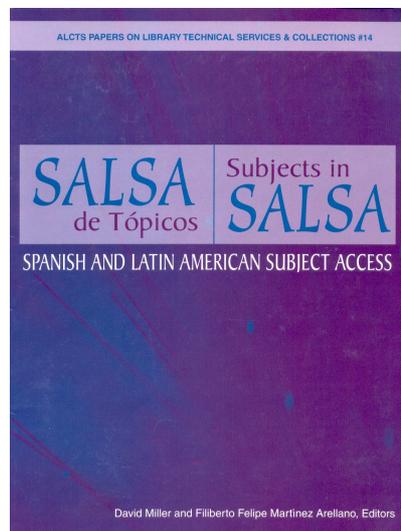
IFLA, August 13, 2008   M. Kreyche - 12

Information about Spanish language subject headings, summarized very briefly here, would have been difficult to compile without the publication of papers from a number of conference held within the last four years. First there was a program at the annual conference of the American Library Association in 2004. The papers appeared in print last year in the Associations first bilingual publication.

Conferences in Latin America

- IFLA, Buenos Aires, 2004
- Encuentros Internacionales
  - Lima, Peru, 2005
  - Mexico, D.F., 2006
  - Buenos Aires, Argentina, 2007
  - Bogota, Colombia, October 2008

A couple months later in 2004, IFLA met in Buenos Aires and Dr. Martínez Arrellano contributed an important paper on subject headings in Mexico. Since then he and others have worked to organize a series of cataloging conferences in Latin America each year; Lima in 2005, Mexico in 2006, Buenos Aires last year, and later this year in Bogota.

The Future of LCSH

- Calhoun, Karen. *The Changing Nature of the Catalog and its Integration with Other Discovery Tools.* (March 2006)
- Task Force on Non-English Access. *Report* (September 2006)
- *On the Record: Report of The Library of Congress Working Group on the Future of Bibliographic Control* (January 2008)

Meanwhile, the last couple of years has seen quite a bit of debate over the future of subject headings, in particular the Library of Congress system. LCSH has been the subject of much controversy over the years, but in March of 2006 a report commissioned by the Library of Congress made a startling recommendation: "Abandon the attempt to do comprehensive subject analysis manually with LCSH in favor of subject keywords; urge LC to dismantle LCSH." This was one of many suggestions listed under the heading of "Innovate and Reduce Costs." Appearing later in 2006, a task force on Non-English Access of the Association for Library Collections and Technical Services did not explicitly take sides on this issues but discussed the importance of subject headings in languages other than English and seemed to assume that LCSH would continue. For now, at least, it seems that this will be the case. The report of the Library of Congress Working Group on the Future, released at the beginning of this year, maintains that "Subject analysis—including analyzing content and creating and applying subject headings and classification numbers—is a core function of cataloging; although expensive, it is nonetheless critical." There also important statements about increasing collaboration on authority work, sharing records, and internationalizing authority files.

Dr. Ann Allan

- Cataloging Professor Emerita (KSU)
- Volunteer work in Costa Rica (1999)

KENT STATE
UNIVERSITY

IFLA, August 13, 2008

M. Kreyche - 15

The seed for the bilingual database was planted as early as 2002 by one of my former cataloging professors, Dr. Ann Allan. She was retired then and had done some volunteer work for libraries in Costa Rica, where she saw the need for better Spanish subject heading tools. In 2003 I spent the year working in El Salvador and saw this need myself at first hand. Meanwhile, Dr. Allan had spent much time making contacts and seeking support for a new Spanish language version of the Library of Congress Subject Headings. I was a little skeptical of such a vast undertaking, but in 2005 I decided to see what I could do in a small way. I hoped to create something that provide some immediate benefits; large projects can take a long time to get started.

Initial goals

- Comprehensive
  - As much data as possible
  - From as many sources as possible
- Efficient and easy to use
  - for technically advanced libraries
  - for libraries with limited resources
- Freely available on the web

KENT STATE
U N I V E R S I T Y

IFLA, August 13, 2008

M. Kreyche - 16

As I saw it, the key problem was to find data and bring as much of it together in one place as possible. It may never be possible to have a single Spanish authority file for the entire Spanish-speaking world, and for the time being, at least, it will be necessary to consult multiple sources of headings. Secondly, the data needed to be presented in a way that was easy and efficient to use. Web technology seemed the best way to accomplish this. And, like other start-up projects on the web, it needed to be freely available so people would use it.

Initial collaborators

- San Francisco Public Library
  - Contribution of bibliographic records
  - Spanish and English headings matched
  - Created a bilingual dictionary of headings
- Queens Borough Public Library
  - Contribution of partially completed file
  - Missing data supplied from San Francisco
  - Completed file merged with San Francisco

The initial sources of data were the San Francisco Public Library and the Queens Library. As I mentioned before, the San Francisco headings exist only in their bibliographic records. The library provided me with a copy of about 16,000 records with Spanish headings and I wrote a computer program to match Spanish headings to the corresponding LC headings in the same record. This allowed me to create a kind of bilingual dictionary. About the time I was finishing this work, the Queens library sent me a copy of their file, which was also a sort of dictionary. Many of the English headings and subdivisions still had no Spanish equivalents, so I wrote another program to look up these terms in the San Francisco file. When there was a match, I copied the Spanish equivalent to the Queens file. This saved quite a bit of work for the Queens staff. After processing these two files I had a bilingual database of nearly 11,000 terms. About a third of the terms were shared by both libraries. Exactly a year after starting the project, I presented the database in public for the first time in Mexico City at the II Encuentro Internacional de Catalogación.

More Data

- Library of Congress Subject Headings
  - Simon Spero (2006)
- Consejo Superior de Investigaciones Científicas (CSIC)
  - CD-ROM (2001)
- Biblioteca Nacional de España
  - Downloaded from authority file
- Bilindex 1986
  - English-Spanish index scanned

Over the next year I obtained data from other sources. Simon Spero, a student at the University of North Carolina, wrote a script to systematically search the Library of Congress authority file and download all its records. He made this data available to the public in December of 2006 and I regard it as a magnificent Christmas gift to the library community. I was able to use the subject records to validate and correct many outdated English terms in the database. I had already asked the Library of Congress for a copy of its machine readable subject authority file but was told that the library is obligated by law to sell this and other products to recover costs. I didn't have the resources to purchase the file, so I am very grateful for Simon's contribution.

Another source I learned about was the CD-ROM published in 2001 by the CSIC in Spain. It was very inexpensive, about $30, so I purchased it and received permission to use contents in the database. Unfortunately, the data is not in MARC format, so it took quite a bit of work to process. The English headings also had a fair number of typographical errors which were very tedious to correct.

I also wrote a script to search the authority file of the Biblioteca Nacional de España and download its subject authority records, much as Simon Spero had done with the Library of Congress records. I had already been in correspondence with the Biblioteca Nacional, which did not have any objections to my use of the data.

Finally, the original printed edition of Bilindex has a very useful English-Spanish index which I scanned and converted to machine readable format.

Current Number of Terms

	Individual	Unique
SFPL	6,618	2,092
Queens	11,134	5,875
CSIC	25,095	16,258
Bilindex	11,952	5,726
BNE	17,758	9,894
Shared by 2 or more		28,393
TOTAL	72,557	52,694

KENT STATE
UNIVERSITY

IFLA, August 13, 2008

M. Kreyche - 19

The size of the database has remained static for the past year, but it seems to have reached a critical mass that makes it useful to a large number of libraries. Even libraries that have subscriptions to commercial sources of subject headings use it. Even though the data is incomplete, outdated, and inconsistent, having it all in one place is very helpful.

Mike's Top 40 libraries for 2007

(100 or more pages served)

Institution/Network	Country	Pages
UA Nuevo León	MX	3027
Verizon, Philadelphia	USA-PA	2947
ITESM	MX	2891
ITESM Monterrey	MX	2654
Newark PL	USA-NJ	2537
RedIRIS	ES	2140
Bellsouth, Miami	USA-FL	2094
Biblioteca de Catalunya	ES	1785
Universitat Politecnica de Catalunya	ES	1683
Queens Library	USA-NY	1371
Southern Light, Mobile	USA-AL	1275
Illinois Century Network	USA-IL	1051
Qwest	USA	896
UNAM	MX	825
Universitat de Vic	ES	655
Universitat de Barcelona	ES	613
rima-tde.net static	ES	523
City of Chula Vista	USA-CA	519
Sertram Networks, Barcelona	ES	496
San Antonio PL	USA-TX	446
San Joaquin Valley Library System	USA-CA	429
UA Barcelona	ES	427
Universitat Pompeu Fabra	ES	391
Oregon State System of Higher Ed	USA-OR	365
Chicago PL	US-IL	338
Kansas City, Kansas PL	USA-KS	328
rima-tde.net	ES	322
Generalitat de Catalunya	ES	299
Universidad Pablo de Olavide	ES	275
Customer-201 uninet-ide.com.mx	MX	262
Multnomah County Library	USA-OR	246
Universitat Oberta de Catalunya	ES	242
ono.com	ES	187
CSIC	ES	177
Phoenix Public Library	USA-AZ	173

10,000 Mexico
10,000 Spain
15,000 U.S.A.
35,000 Total

KENT STATE
UNIVERSITY

IFLA, August 13, 2008   M. Kreyche - 20

Statistics from the web server logs show that approximate 35,000 pages were served during the year 2007. These were divided more or less evenly between Spain, Mexico, and the United States, with U.S. libraries registering a slightly higher number.

National Endowment for theHumanities

- $30,000 grant
- MARC format
- Greater convenience and personalization
- Interactive corrections and comments
- Community-based collaboration
- Web services (machine-to-machine)
- Batch processing

KENT STATE
U N I V E R S I T Y

IFLA, August 13, 2008   M. Kreyche - 21

A year ago I received funding from the National Endowment for the Humanities to continue work on the project. Most of the money is being used to hire a programmer to add functionality to the database. The major achievements to date have been the conversion of all the data to MARC format, redesigning of the database to accommodate the MARC structure and the Unicode character set, partial integration of the Library of Congress subject authority file, and a login system.

The login system will support personalized configuration options and interactive contributions in the form of corrections, suggestions, and comments. The idea is to form a collaborative, international community of catalogers who use Spanish language subject headings. It is also important to think in terms of automated processes like batch processing of MARC bibliographic records and web services to facilitate machine-to-machine interactions. In the coming months I will be soliciting suggestions along these lines.

Now that the MARC structure is in place I will renew my attempts to acquire more data and update existing records where necessary.

Next, I would like to give a brief demonstration of the system.

The Current Database Interface
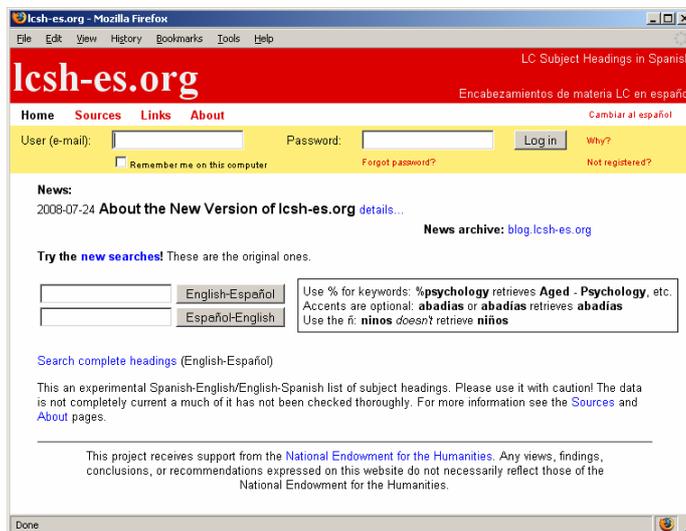
The current version has been available since the middle of May. The home page of the provides the original search interface to the earlier, non-MARC data. So far most people are continuing to use it instead of logging in to use the new interface. This will change soon; everyone will be required to log in and the old interface will have less prominence and will eventually be removed.

Notice also the "News" section, which is really a feed from a blog. Actually, there are two blogs, one in Spanish and one in English, corresponding to the bilingual interface of the database. The language that appears by default depends on the language of the browser, but it's easy to change languages with the link near the upper right corner.

Search Page

Let's log in and have a look at the search page. Currently there are four searches: English (to Spanish) and Spanish to English against the bilingual database, and a heading search and numeric search against the Library of Congress file.

English Search

Home : Search : Results

8 terms for search: **%canada**   Repeat or modify: %canada   English-Español

< Previous Next >

|			English	Español
SF QB	1	– Canada	– Canadá
BX	QB	2	Canada	Canadá
NE		3	Canada - History - To 1763 (New France)	Canadá - Historia - Hasta 1763
NE		4	Canada - Literatures	Literatura canadiense
NE CS		5	Indians of North America - Canada	Indios de Canadá
BX		6	Indians of North America - Canada	Indios del Canadá
NE		7	Lakes - Canada	Lagos - Canadá
NE CS		8	Mountains - Canada	Montañas - Canadá

Key
BX Bilindex (1984)   SF San Francisco Public Library   QB Queens Borough Public Library
NE Biblioteca Nacional de España
CS Consejo Superior de Investigaciones Científicas-España (2000)

This project receives support from the National Endowment for the Humanities. Any views, findings, conclusions, or recommendations expressed on this website do not necessarily reflect those of the National Endowment for the Humanities.

Done

KENT STATE
U N I V E R S I T Y

IFLA, August 13, 2008   M. Kreyche - 24

Let start with a simple English search for "Canada". Searches are left anchored by default, but using the percent character (%) as a prefix retrieves the string in any position; so "Indians of North America—Canada" and "Lakes—Canada" and "Mountains—Canada" are now included. Clicking on the icon next to the English heading displays the LC authority record. If there is no icon, the heading has not been verified against the LC file.

LC Authority Record
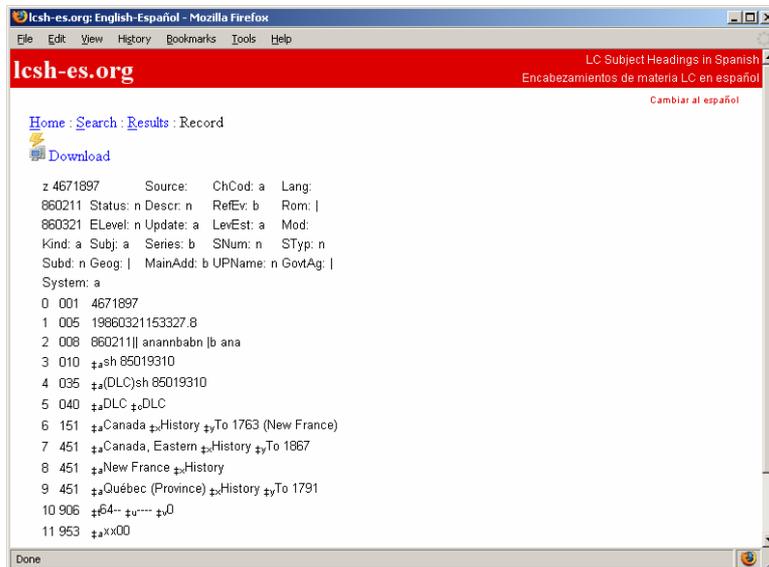
Note there is a download link. This feature is still under development and there are some character set problems that remain to be solved, but it's possible to import these records into a local system.

Authority Records

Let's go back to the search page. The abbreviations in the left hand columns show the sources of each heading. If the heading is shared by more than one source, all of them are indicated.Clicking on the abbreviation shows the record for that source. Let's look at a record from the Biblioteca Nacional de España because that's the only source that has full records.

Spanish Authority Record

Note the 751 field at the bottom of the record. This did not exist in the original record downloaded from the BNE authority file. I reformatted the text in all the LCSH 670 fields and wrote a program to search for the heading in the LC file. Whenever I found a match, I inserted the 1xx field from the LC record as a 7xx field in the Spanish record and added subfield zero with the linking number to the LC record.

Easy Copying and Pasting

Let's go back to search results again and look at a feature that simplifies copying and pasting. If you move the cursor over a Spanish heading, it is shaded.

Highlight with one click
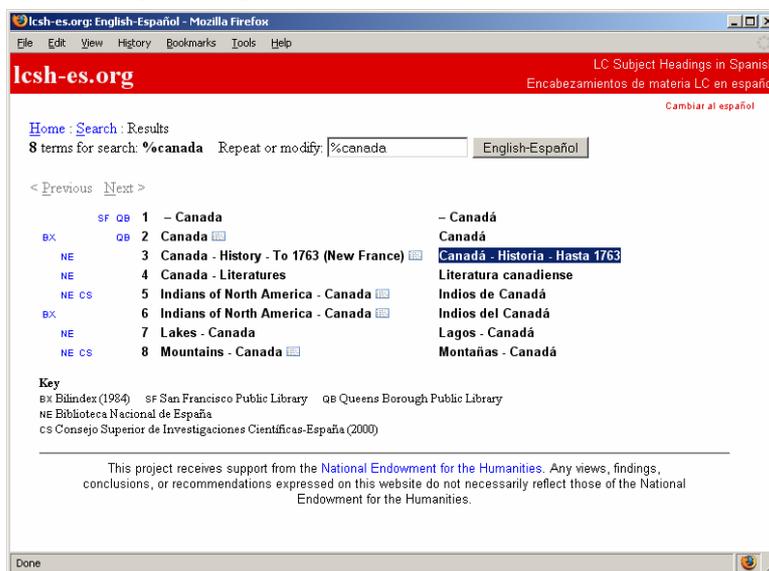
Click the heading and it becomes fully highlighted—no need to drag the mouse!

Ready for Copying and Pasting
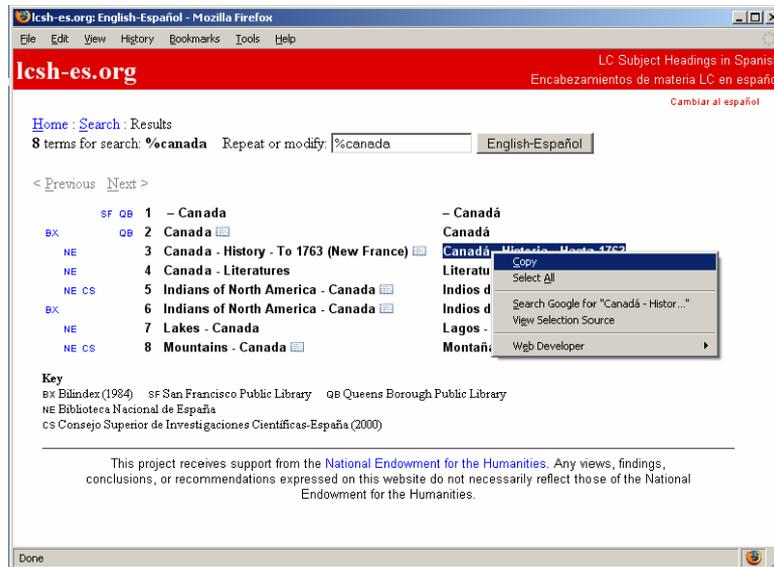
Right click to get the context menu or use Control-C—you're ready to past the heading into another program.

Search Page

Now let's try a Spanish search. Note the option to show "See also" references. Let's select it and search "Canadá".

Spanish Search

Notice that it's not necessary to put the accent on the letter "a". It will match either way. Note also that the letter "n" matches the "ñ" (enye). This is a change from the earlier version of the database. English speakers not accustomed to Spanish keyboard will welcome the change, but the reason for has to do with the character sorting of the underlying database.

Merci beaucoup!

¡Muchas gracias!

Thanks very much!

That's all we have time for now. I'll be glad to take questions!