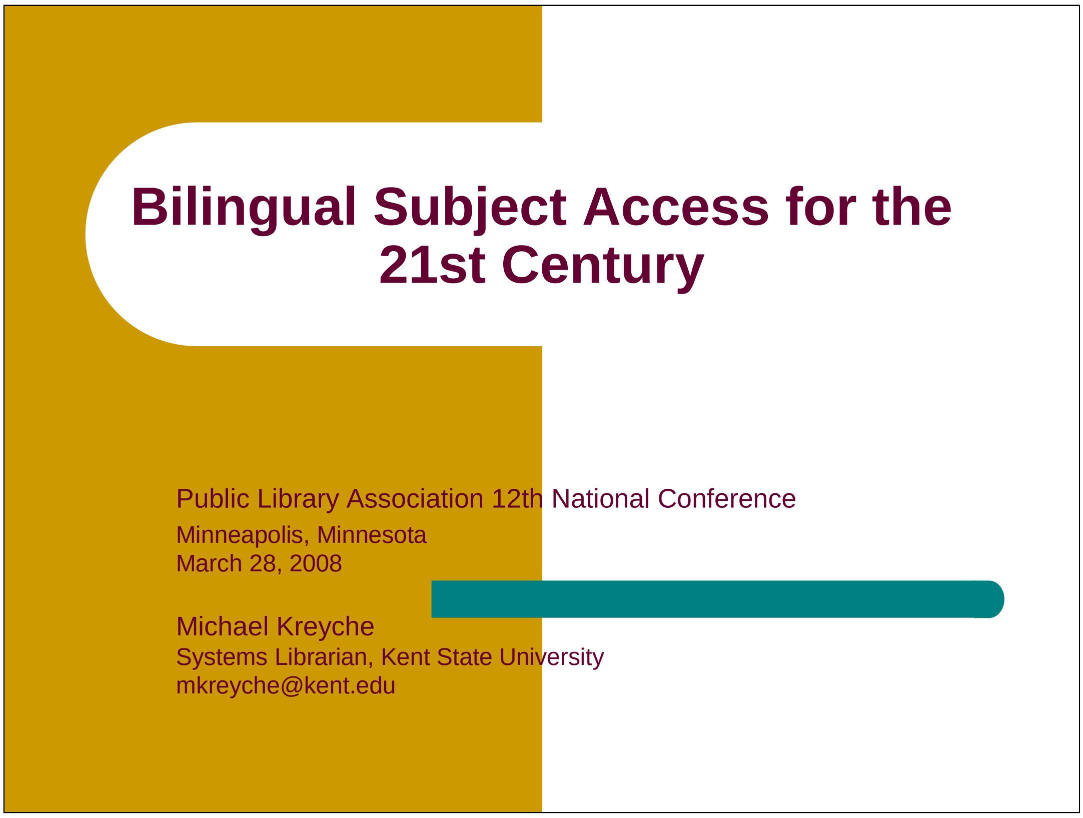
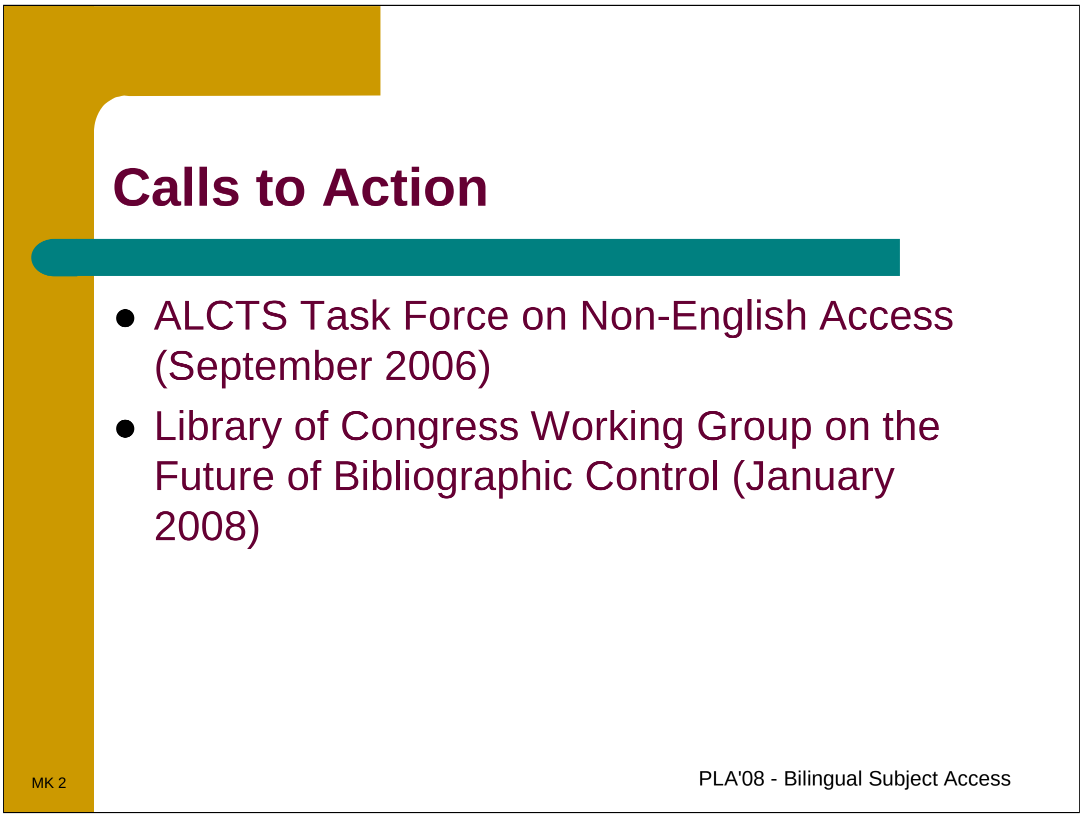


Bilingual Subject Access for the 21st Century

Public Library Association 12th National Conference

Minneapolis, Minnesota
March 28, 2008

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



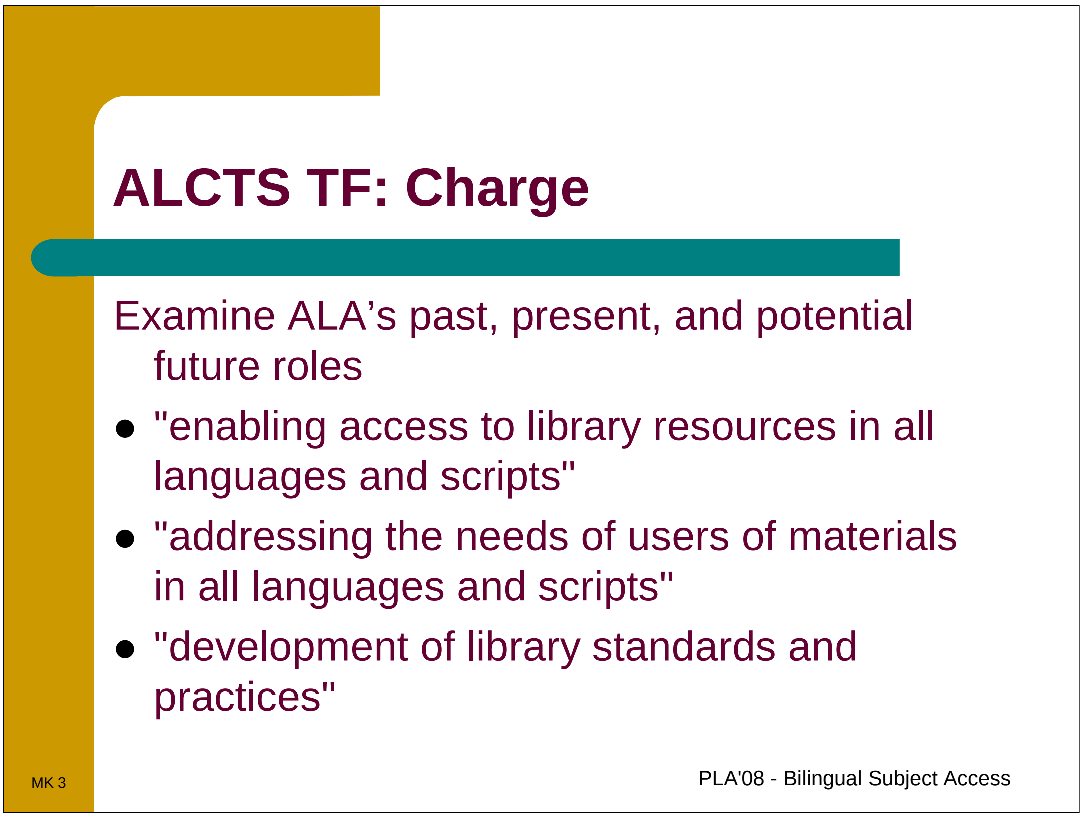
Calls to Action

- ALCTS Task Force on Non-English Access (September 2006)
- Library of Congress Working Group on the Future of Bibliographic Control (January 2008)

MK 2

PLA'08 - Bilingual Subject Access

Bilingual or multilingual subject access isn’t a new idea; some people have been working on it for years. What’s different now is the wider attention it’s getting. I’m thinking in particular of two reports released within the last year and a half. The first is that of the ALCTS Task Force on Non-English Access of September 2006 and revised about a year ago. The second is that of the LC Working Group on the Future of Bibliographic Control. Allow me to highlight a few items from these reports.

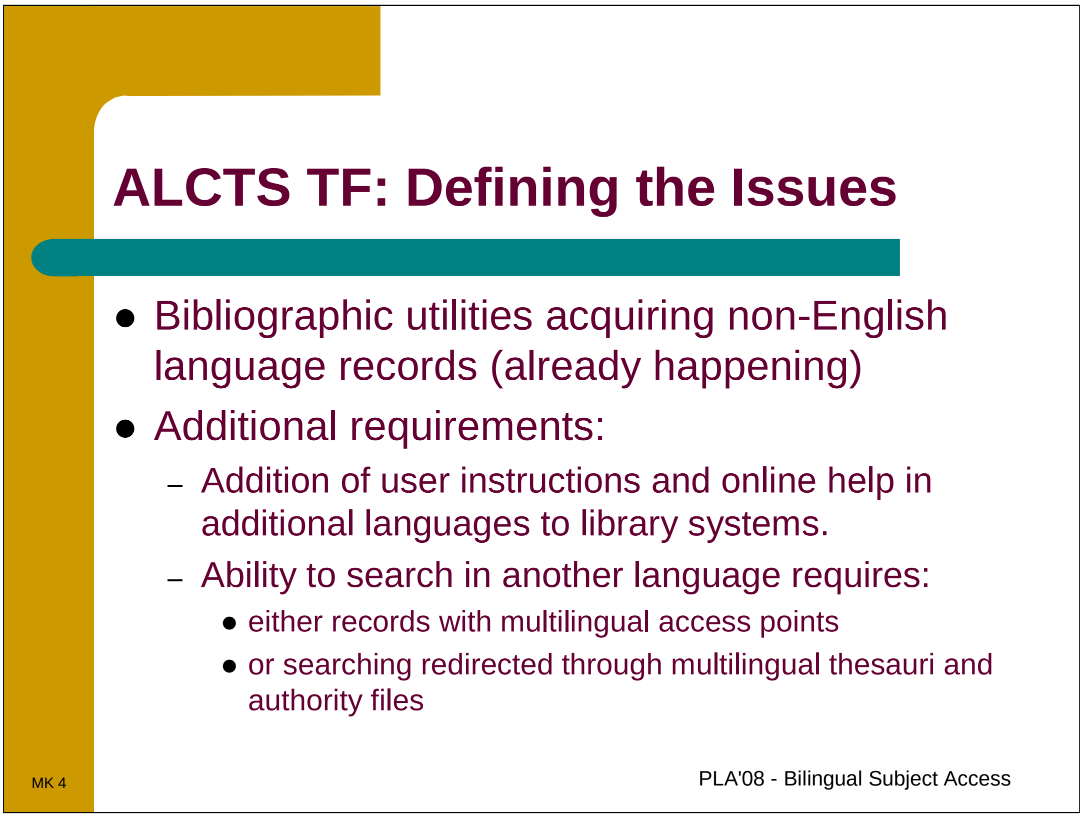


ALCTS TF: Charge

Examine ALA’s past, present, and potential future roles

- "enabling access to library resources in all languages and scripts"
- "addressing the needs of users of materials in all languages and scripts"
- "development of library standards and practices"

First of all, the charge to the ALCTS Task Force was quite broad: examining ALA’s role in addressing issues regarding materials in all languages and scripts. Everything I have to say today concerns only Spanish and English, but I believe there are implications for other languages as well. The report is quite detailed and I’m going to mention just a few points from the executive summary.



ALCTS TF: Defining the Issues

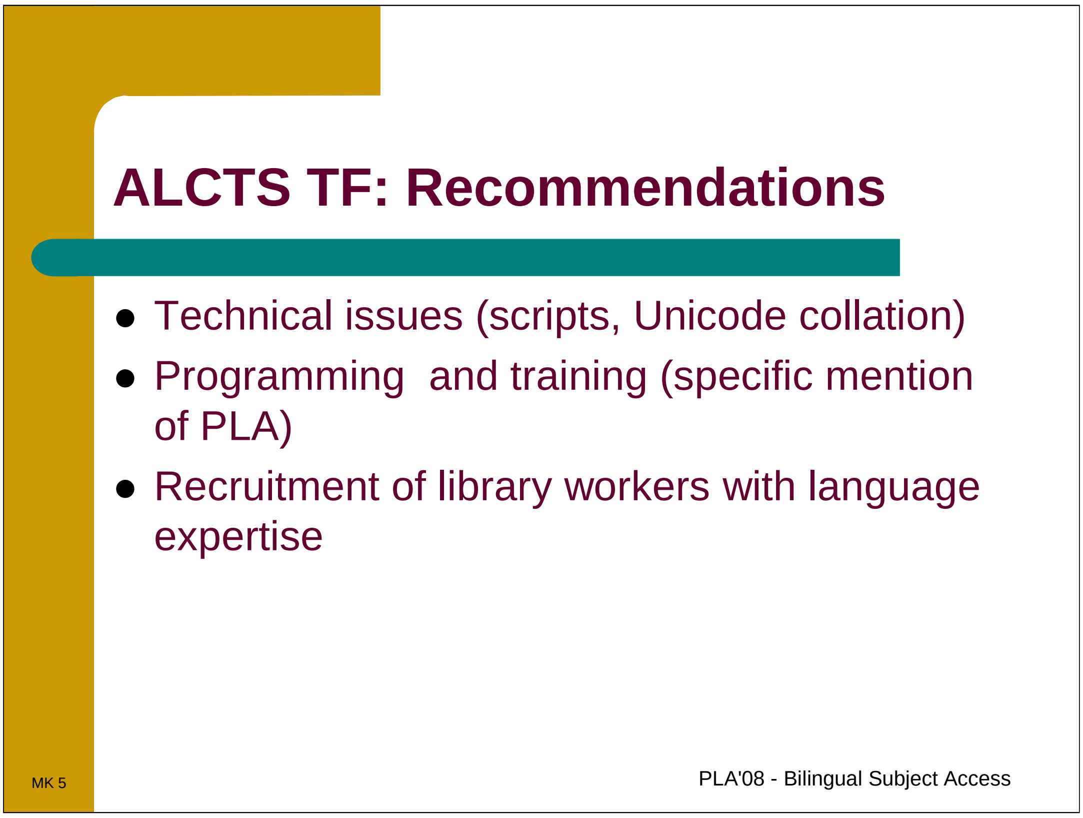
- Bibliographic utilities acquiring non-English language records (already happening)
- Additional requirements:
  - Addition of user instructions and online help in additional languages to library systems.
  - Ability to search in another language requires:
    - either records with multilingual access points
    - or searching redirected through multilingual thesauri and authority files

MK 4

PLA'08 - Bilingual Subject Access

One of observations in the report is that the bibliographic utilities are acquiring non-English language records from book vendors. I know from scanning messages on the OCLC-CAT email list that OCLC is also batch loading records from national libraries in other countries, but I don't know if these include countries where Spanish is spoken. The result is the creation of so-called "parallel records" for the same title. One is created in English, by an English-speaking cataloger; the other in another language, by a speaker of that language, generally the language of the work being cataloged. The obvious advantage of using the latter is that our catalogs can show the eventual readers, our patrons, the bibliographic descriptions in their own language.

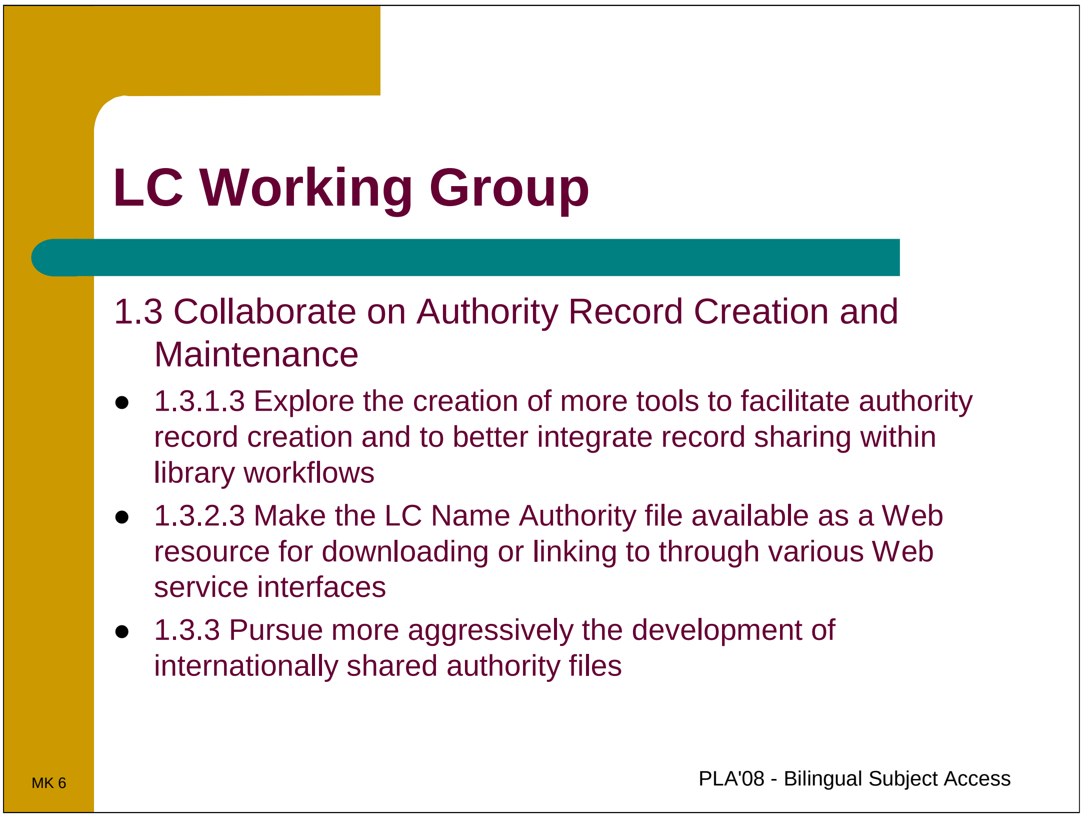
The Task Force also points out a couple other requirements for libraries that serve non-English speaking populations. The first is user instruction and online help in alternate languages. This seems fairly common on library web sites and in online catalogs. The other is the ability to search in other languages, a somewhat more difficult proposition, and the one that I'm here to talk about.



ALCTS TF: Recommendations

- Technical issues (scripts, Unicode collation)
- Programming and training (specific mention of PLA)
- Recruitment of library workers with language expertise

To wind up my comments on the ALCTS Task Force’s report, I’ll mention just a few of the recommendations. First there are a host of technical issues that I’m lumping into a single point. Fortunately, Spanish doesn’t require a completely different script from English, but the techniques the library community developed for representing diacritics and special characters—long before anybody else was doing it—are not suitable for the modern computing environment. The next point is the need for training and programming; and today’s panel is an excellent example of putting this into practice. In fact, several recommendations specifically mention working with PLA. The third is the need to recruit library workers with language expertise.



LC Working Group

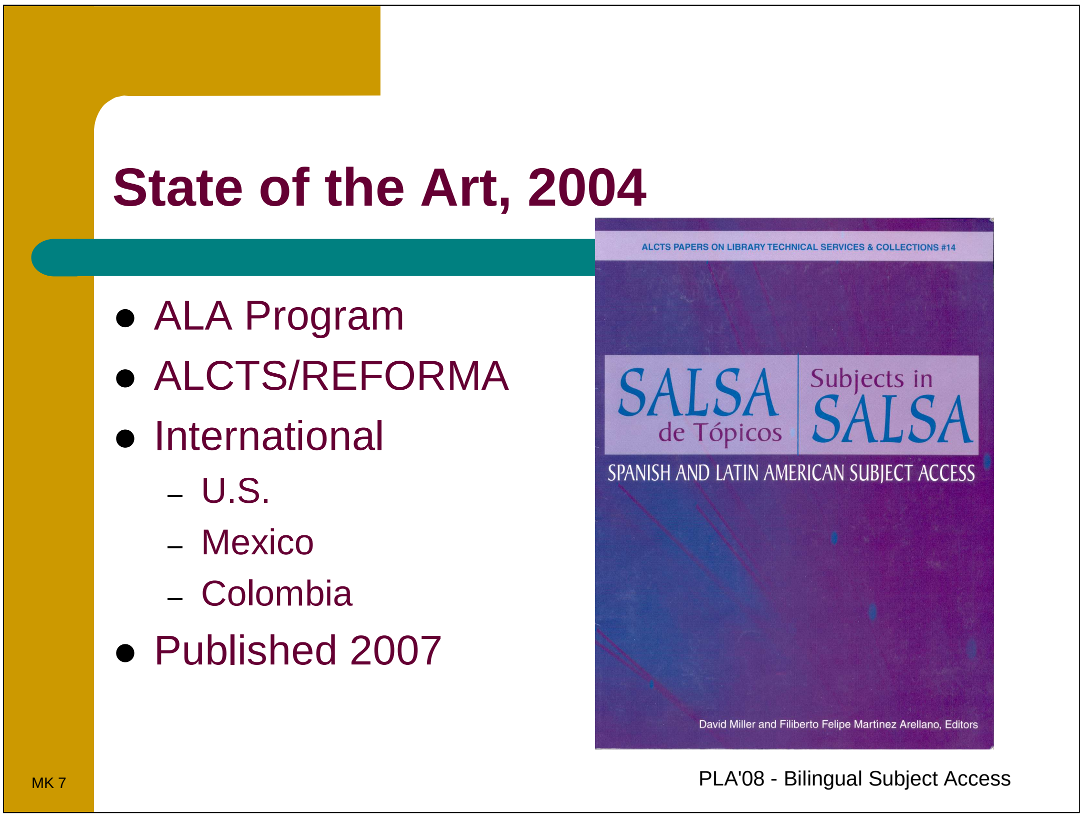
1.3 Collaborate on Authority Record Creation and Maintenance

- 1.3.1.3 Explore the creation of more tools to facilitate authority record creation and to better integrate record sharing within library workflows
- 1.3.2.3 Make the LC Name Authority file available as a Web resource for downloading or linking to through various Web service interfaces
- 1.3.3 Pursue more aggressively the development of internationally shared authority files

MK 6

PLA'08 - Bilingual Subject Access

The LC Working Group on the Future of Bibliographic Control certainly got a lot more press than the ALCTS Task Force and I'm sure some of you are more familiar with it than I am, but I'm going to mention a few recommendations from the section on Authority Records. The first one that jumped out at me was 1.3.1.3: “Explore the creation of more tools to facilitate authority record creation and to better integrate record sharing in library workflows.” This is exactly what I've been working towards, a new kind of tool, for the last two and a half years, so I felt a real sense of validation when I read this. Another, more specific recommendation was, “Make the LC Name Authority file available as a web resource, for downloading or linking to through various web services.” The first question this raised in my mind was, “What about the Subject Authority file, too?” Here again, some of us have been working on this already, thanks to an enterprising young man, Simon Spero, at the University of North Carolina. In the fall of 2006 he wrote a script to systematically download LC's authority records and he's made the raw data available to others. His own interest is analyzing the subject authority file and restructuring it into a true thesaurus, but at least one other person besides myself has used the records to build an online database of subject headings that's much easier to use than LC's. Mine isn't public yet, but I'll give you a peek in a few minutes.



State of the Art, 2004

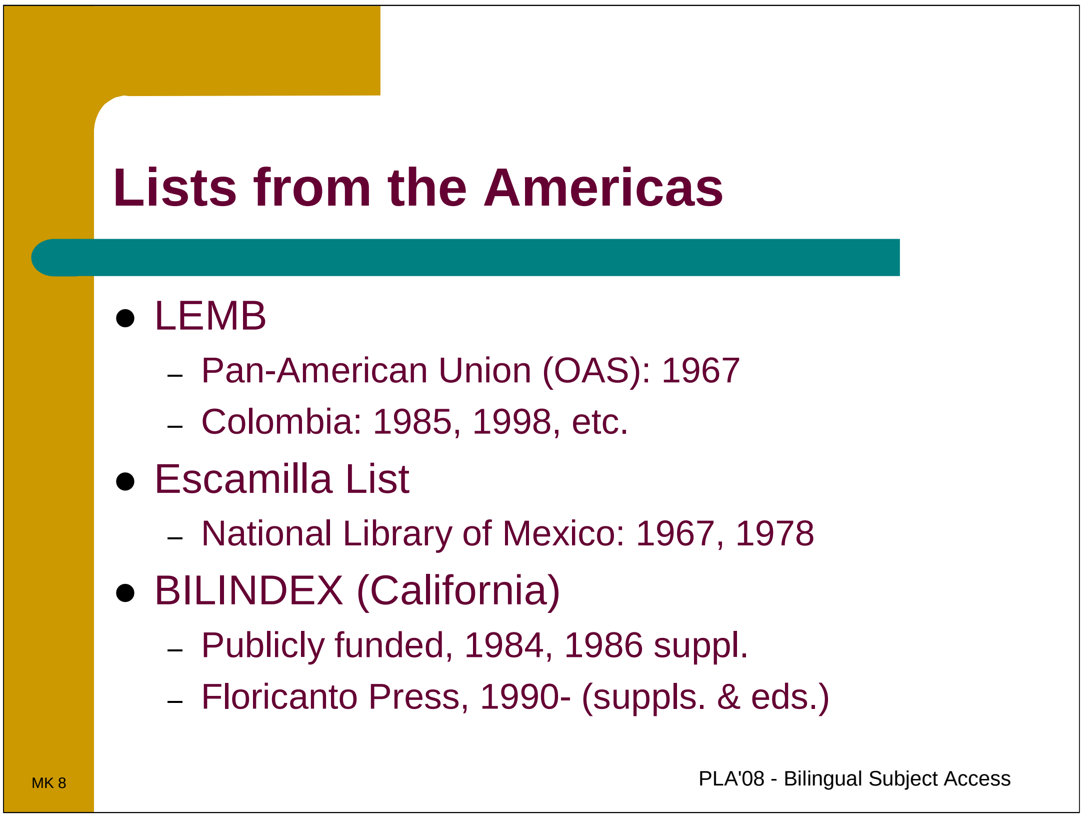
- ALA Program
- ALCTS/REFORMA
- International
  - U.S.
  - Mexico
  - Colombia
- Published 2007

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 Martínez Arellano, Editors

MK 7

PLA'08 - Bilingual Subject Access

Another recently published resource that I highly recommend is “SALSA de Tópicos/Subjects in SALSA,” the first bilingual book that ALA has published. It’s based on a program sponsored by ALCTS and REFORMA at the 2004 ALA Conference in Orlando. It’s required reading for anyone interested in Spanish subject headings and David Miller of Curry College deserves a huge amount of credit for editing this book and seeing it into print. As soon as you’re back at work on Monday, have your library order a copy.



Lists from the Americas

- LEMB
  - Pan-American Union (OAS): 1967
  - Colombia: 1985, 1998, etc.
- Escamilla List
  - National Library of Mexico: 1967, 1978
- BILINDEX (California)
  - Publicly funded, 1984, 1986 suppl.
  - Floricanto Press, 1990- (suppls. & eds.)

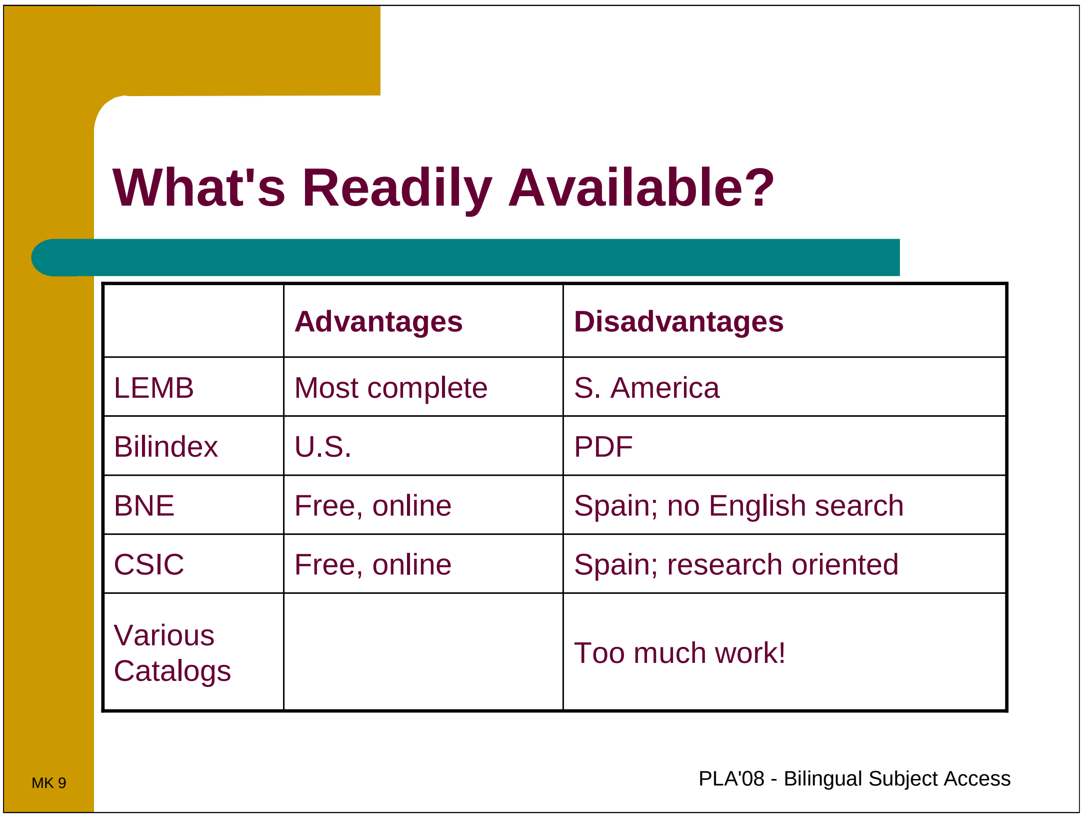
MK 8

PLA'08 - Bilingual Subject Access

Let me distill some of the information you can find in SALSA by describing the major sources of subject headings in the Americas. These aren't the earliest or the only lists, but they're the most influential. The LEMB, Lista de Encabezamientos de Materia para Bibliotecas, is the most complete and got its start in 1960s under the auspices of the Panamerican Union, now know as the Organization of American States. The list has passed through several phases of sponsorship and is centered in Colombia. It's available fby subscription either online or on CD-ROM, but doesn't seem to be heavily marketed in North America.

The list created by Gloria Escamilla from the National Library of Mexico is more or less the standard in that country, but it's very outdated and I'm not sure to what extent it's being maintained internally within the National Library. The libraries in Mexico most serious about authority work value the Escamilla list highly, but rely heavily on the LEMB for its currency and comprehensiveness. An early form of the Escamilla list was contributed to the original editors of the LEMB; notice that both were first published in the same year, 1967.

Bilindex is probably the most common source of Spanish headings in North America. It was originally created with public funding, and when that support ended, it spun off into a commercial venture under the imprint of Floricanto Press.



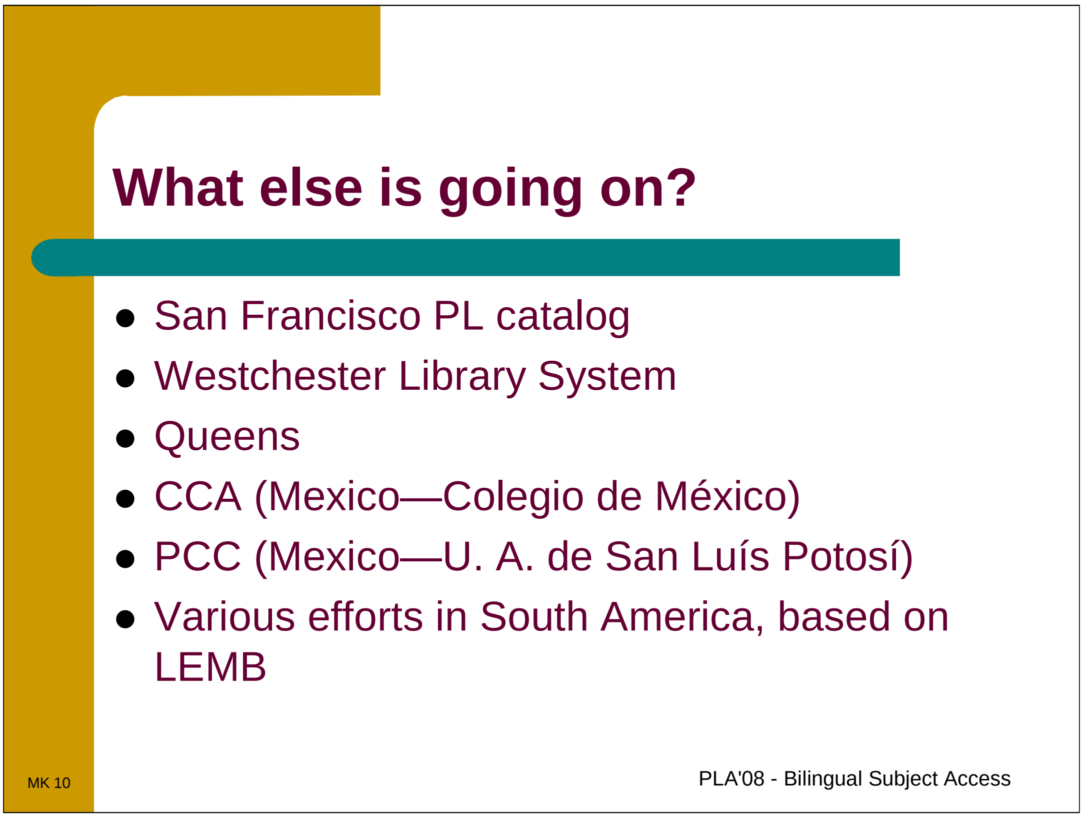
What's Readily Available?

	Advantages	Disadvantages
LEMB	Most complete	S. America
Bilindex	U.S.	PDF
BNE	Free, online	Spain; no English search
CSIC	Free, online	Spain; research oriented
Various Catalogs		Too much work!

MK 9

PLA'08 - Bilingual Subject Access

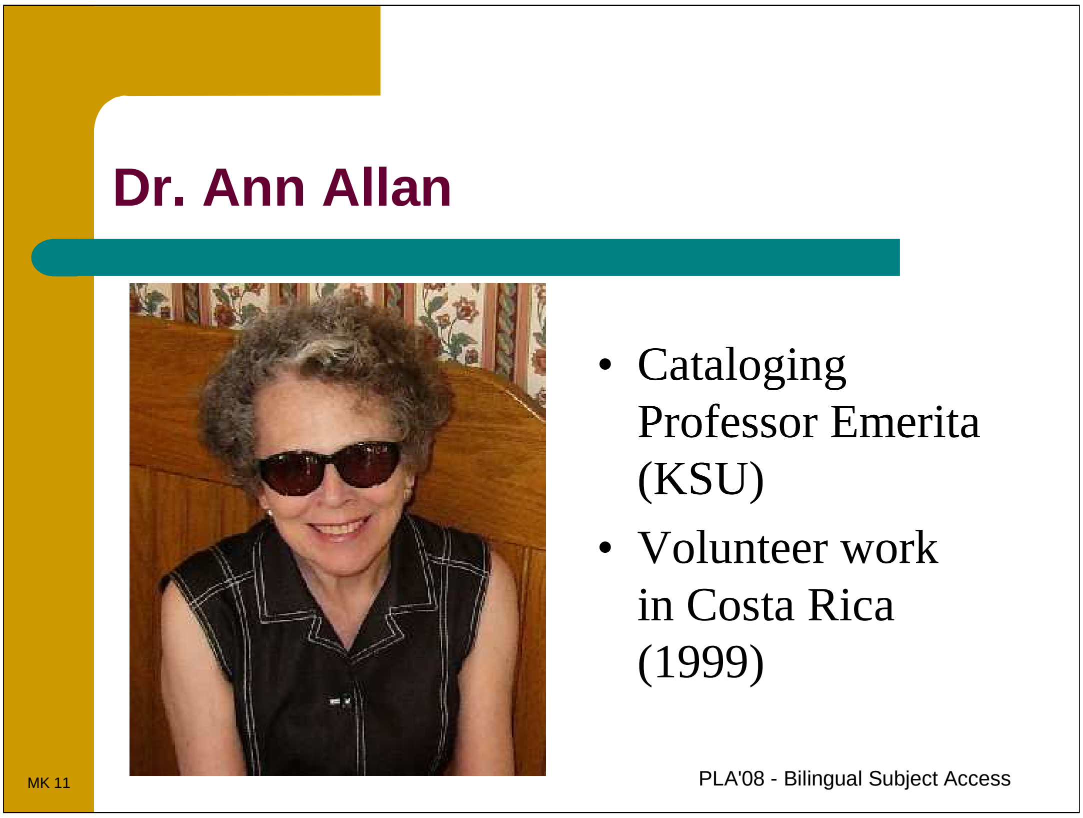
The next slide compares these three sources along with a couple from Spain. The LEMB is the most complete, and its South American bias is somewhat of a disadvantage for libraries in the northern half of our hemisphere where Mexican and Caribbean influences predominate. I'm told it includes English terms from LCSH, but they're not indexed. The source records are in MARC format, but the commercial product doesn't feature MARC ouput. The chief disadvantage of Bilindex is that it's in PDF format. Unless it gets converted to MARC at some point, I can't see that is has much of a future. The National Library of Spain makes its authority file available on the web through its catalog as does a cooperative program of research libraries sponsored by a governmental agency, the CSIC or Consejo Superior de Investigaciones Científicas. Both are available on CD-ROM as well. The biggest disadvantage to us in this hemisphere is their European language bias and their relatively small size. Apart from these sources, none of which is ideal, the only option is to look in catalogs to see what other libraries are using, which is extremely time consuming.



What else is going on?

- San Francisco PL catalog
- Westchester Library System
- Queens
- CCA (Mexico—Colegio de México)
- PCC (Mexico—U. A. de San Luís Potosí)
- Various efforts in South America, based on LEMB

There’s no lack of work getting done on Spanish subject headings. The main problem, in my opinion, is that much of it is not being shared, and so far there hasn’t been a good mechanism for sharing. The San Francisco Public Library has a long standing reputation for putting Spanish language headings into its catalog. Westchester Library System in Connecticut was funded a few years ago to create Spanish authority records and use them in their catalog. As we’re hearing first-hand today, Queens Library has undertaken a significant project to convert to build their own list of headings and add them to their catalog. There are a couple of cooperative programs in Mexico, but the cooperation is within the two groups, not between them. And finally, there are several libraries or groups of libraries in South America that are basing their authority work on the LEMB, but as far as I can tell, they’re not feeding their work back to the LEMB.



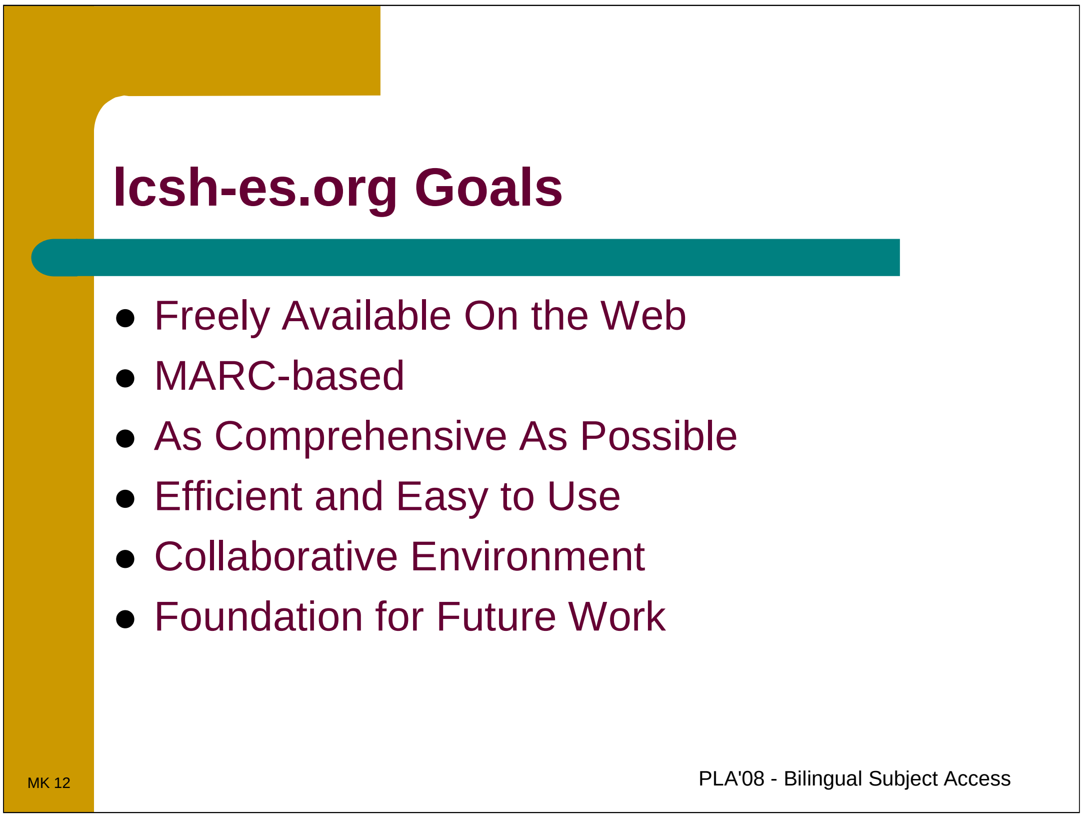
Dr. Ann Allan

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

MK 11

PLA'08 - Bilingual Subject Access

I first got interested in Spanish subject headings through Ann Allan, a former cataloging professor of mine from Kent State. After her retirement she did some volunteer work in Costa Rica where she saw the need for a good subject heading tool. She even motivated me to spend a few hours visiting a library in Mexico while on an anniversary trip with my wife back in 2002. I saw the need myself at first hand when I spent a year in El Salvador with the Fulbright program. And seeing the lack of resources there also made me realize that part of the definition of a “good” tool was “free or low-cost.”



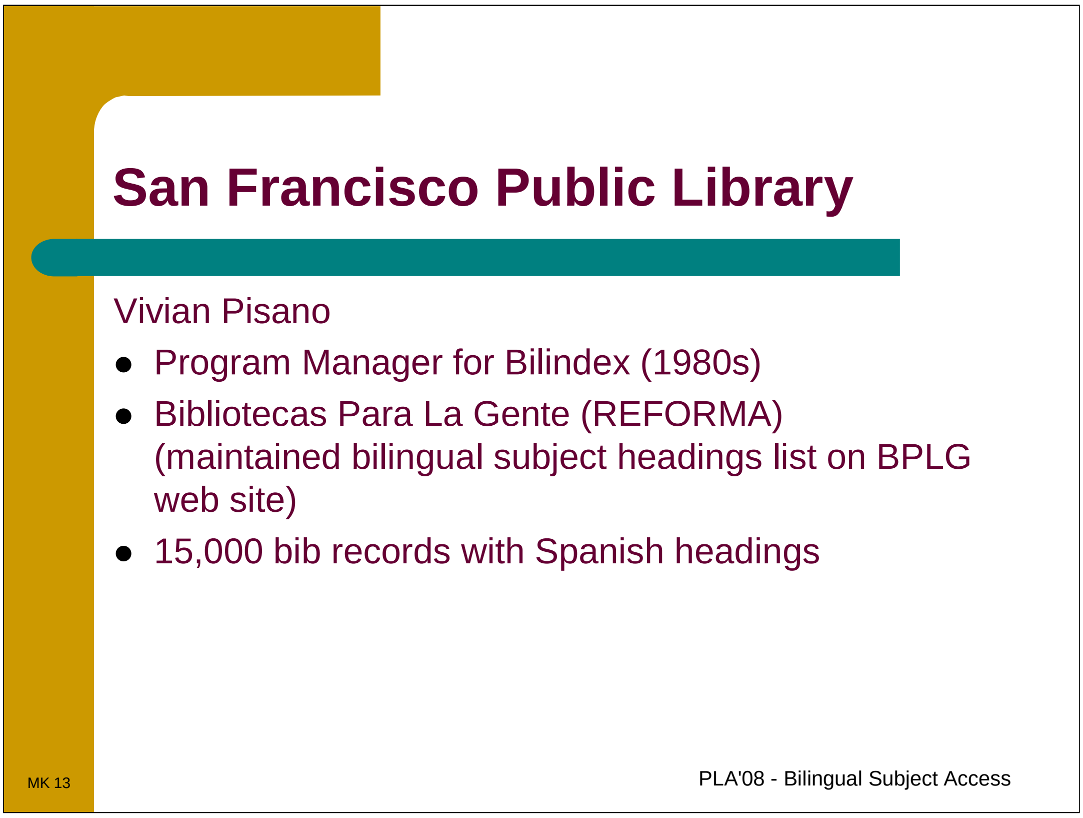
lcsh-es.org Goals

- Freely Available On the Web
- MARC-based
- As Comprehensive As Possible
- Efficient and Easy to Use
- Collaborative Environment
- Foundation for Future Work

MK 12

PLA'08 - Bilingual Subject Access

I started thinking about the problem seriously in 2005 and formulated some goals for the project that I later called lcsh-es.org (es being the international abbreviation for the Spanish language). First of all, I thought a modern subject heading tool should be web-based and freely available. It should also fully support the MARC format and be as comprehensive as possible. It needed to be efficient and easy to use and foster a collaborative environment. And if not all these goals could be sustained, at least it should provide a foundation for future work. Even before I had any data, I started designing a database and, armed with a list of contacts that Ann had accumulated over a period of several years, I made my first call to Vivian Pisano at the San Francisco Public Library.

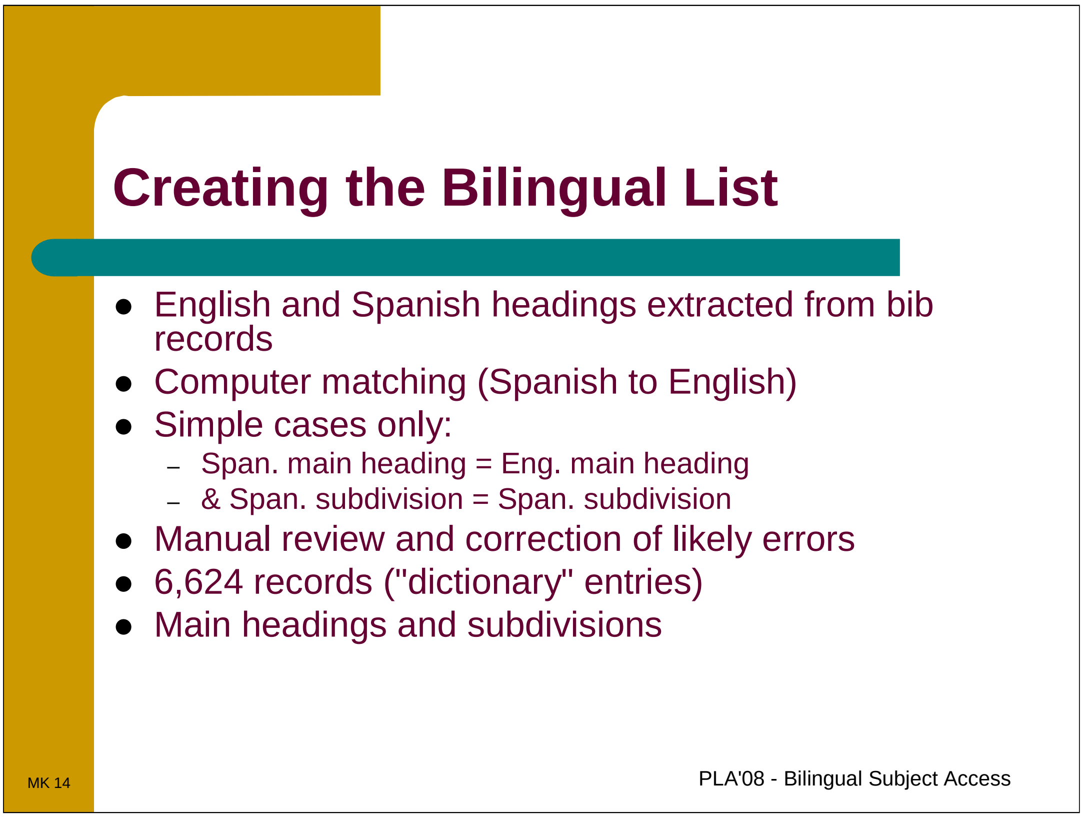


San Francisco Public Library

Vivian Pisano

- Program Manager for Bilindex (1980s)
- Bibliotecas Para La Gente (REFORMA) (maintained bilingual subject headings list on BPLG web site)
- 15,000 bib records with Spanish headings

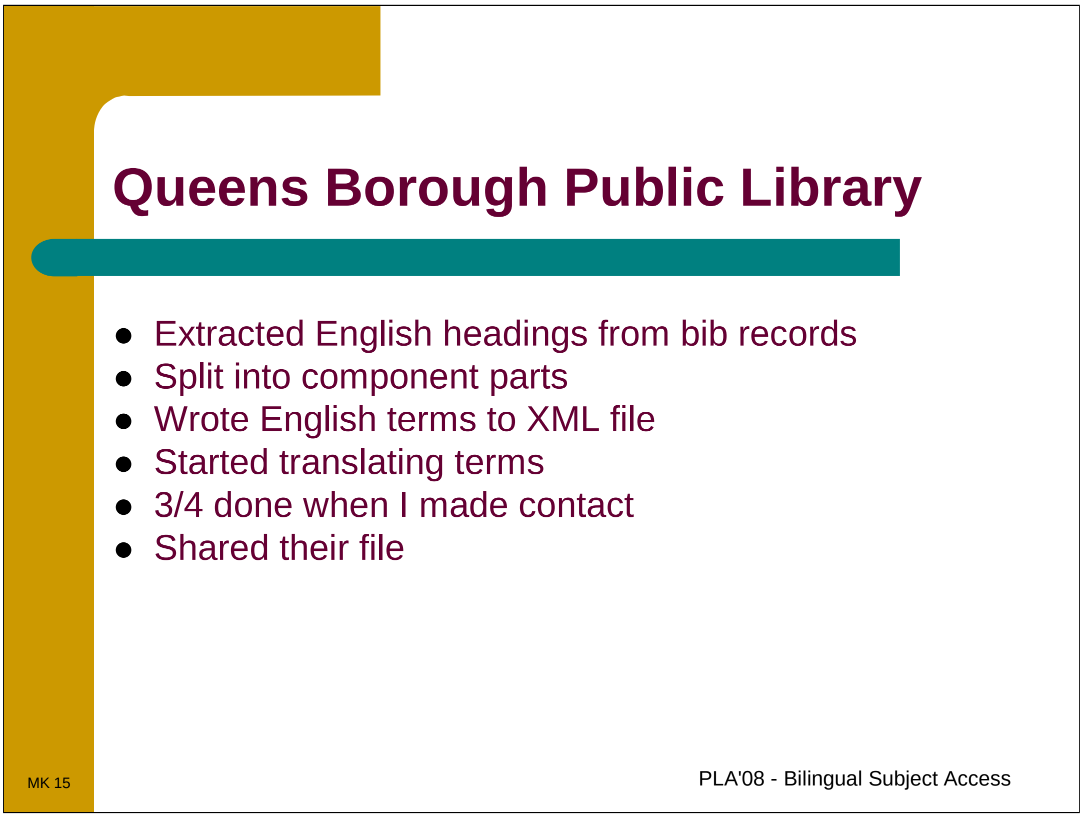
Vivian had worked on the original Bilindex project, was very involved with her local REFORMA chapter. For a number of years she maintained a list of Spanish subject headings on its web site. After a conversation or two, she agreed to supply me with copies of San Francisco’s bib records for Spanish language material. They contained both English and Spanish subject headings, and my idea was to match up the equivalent headings.



Creating the Bilingual List

- English and Spanish headings extracted from bib records
- Computer matching (Spanish to English)
- Simple cases only:
  - Span. main heading = Eng. main heading
  - & Span. subdivision = Span. subdivision
- Manual review and correction of likely errors
- 6,624 records ("dictionary" entries)
- Main headings and subdivisions

You can see here the steps involved in that process—extracting the Spanish and English headings from the records, developing an algorithm for matching individual terms in the two languages, thereby generating a sort of dictionary, and then reviewing the file for errors. I ended up with over 6,600 entries, both main headings and subdivisions.



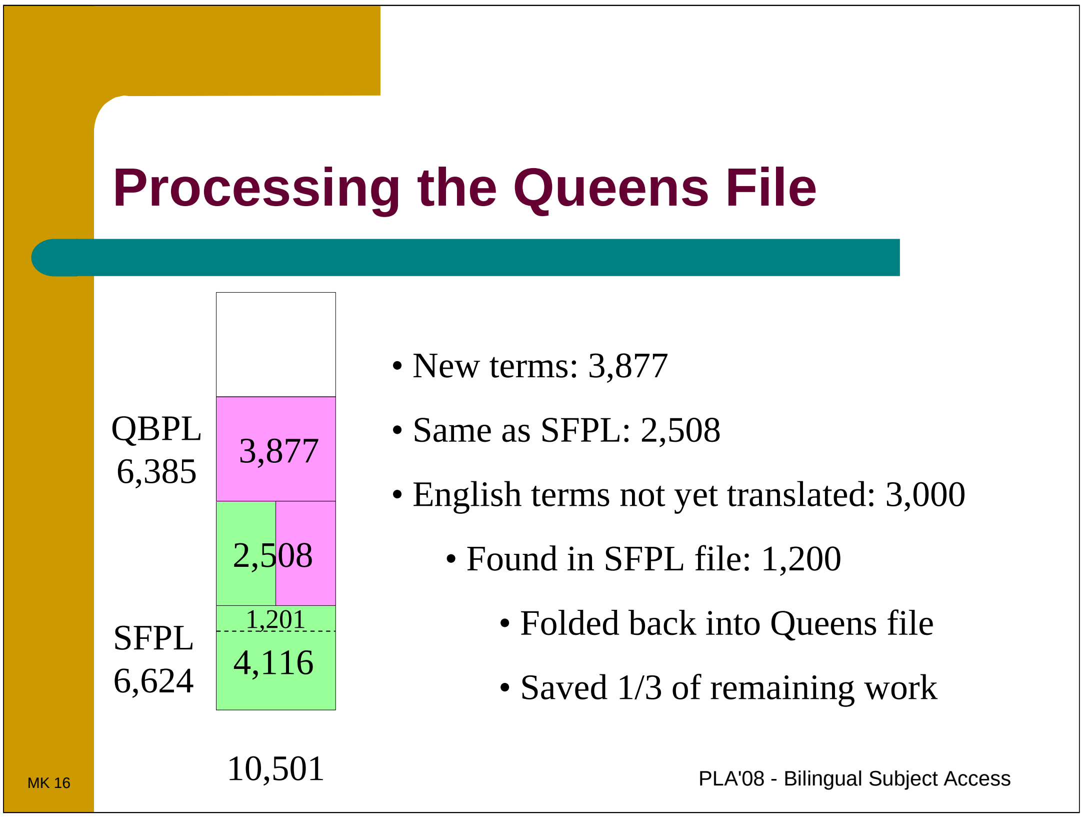
Queens Borough Public Library

- Extracted English headings from bib records
- Split into component parts
- Wrote English terms to XML file
- Started translating terms
- 3/4 done when I made contact
- Shared their file

MK 15

PLA'08 - Bilingual Subject Access

. While working with the San Francisco headings, I made contact with Queens to learn about their project and during a visit Ann made there she secured their agreement to share their work with me. As we'l hear, their process was to extract English headings from their bib records, split them into their component parts, and create a dictionary by establishing the corresponding Spanish terms. There was a great deal of similarity to what I was doing with the San Francisco data—the main difference was that I was matching English terms to Spanish ones that were already there, while Queens only had the English terms and had to find Spanish equivalents. At this point they were about three quarters of the way through that process.

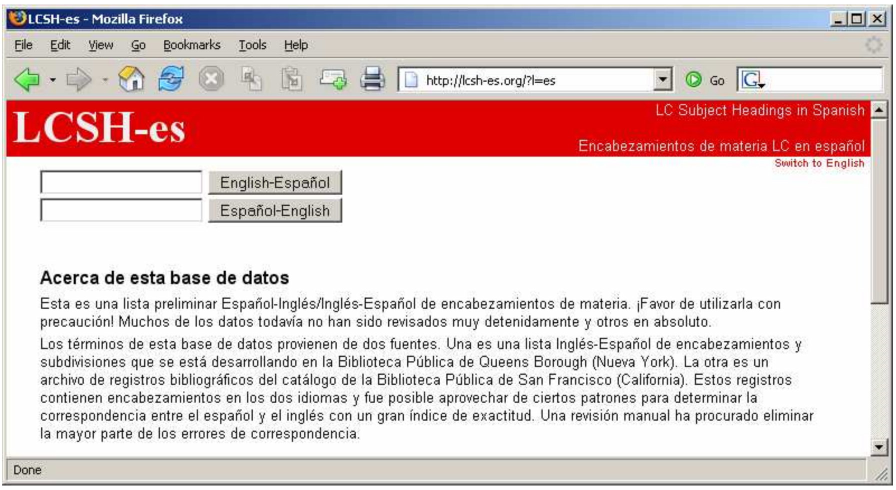


Processing the Queens File

- New terms: 3,877
- Same as SFPL: 2,508
- English terms not yet translated: 3,000
  - Found in SFPL file: 1,200
    - Folded back into Queens file
    - Saved 1/3 of remaining work

I was able to look up in my San Francisco data the English terms from Queens that still needed a translation, and I managed to fill in a good portion of the blanks. I sent the results back to Queens, saving about third of their remaining work. By the summer of 2006 I was able to merge the completed Queens file with the San Francisco data for a total of over 10,000 terms. Roughly a third was unique to each library and the other third shared between them.

lcsh-es.org: Mexico, Sept. '06

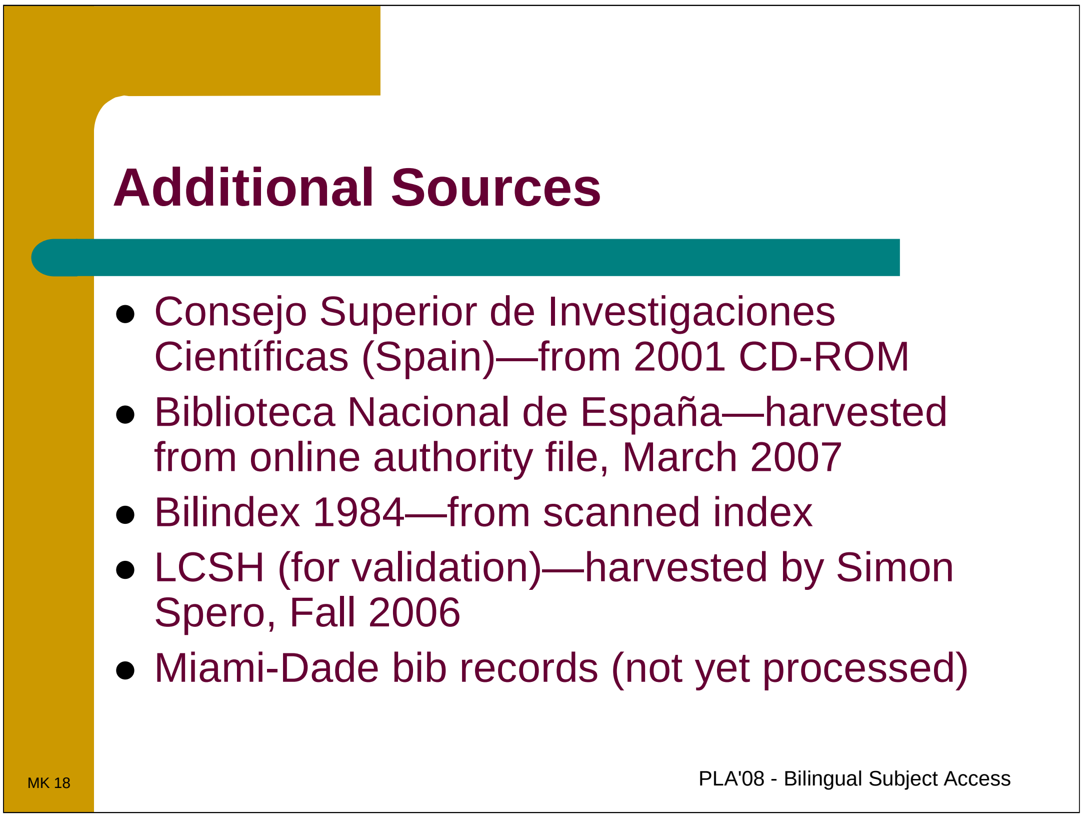

LCSH-es - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://lcsh-es.org/?l=es
Go
LCSH-es
LC Subject Headings in Spanish
Encabezamientos de materia LC en español
Switch to English
English-Español
Español-English
Done

Acerca de esta base de datos

Esta es una lista preliminar Español-Inglés/Inglés-Español de encabezamientos de materia. ¡Favor de utilizarla con precaución! Muchos de los datos todavía no han sido revisados muy detenidamente y otros en absoluto.

Los términos de esta base de datos provienen de dos fuentes. Una es una lista Inglés-Español de encabezamientos y subdivisiones que se está desarrollando en la Biblioteca Pública de Queens Borough (Nueva York). La otra es un archivo de registros bibliográficos del catálogo de la Biblioteca Pública de San Francisco (California). Estos registros contienen encabezamientos en los dos idiomas y fue posible aprovechar de ciertos patrones para determinar la correspondencia entre el español y el inglés con un gran índice de exactitud. Una revisión manual ha procurado eliminar la mayor parte de los errores de correspondencia.

In September of 2006 I made the first public announcement of the database in a paper presented at a conference in Mexico City. As you can see from the page in this slide, I designed the site from the beginning to have a bilingual interface. Even though I only had the bare skeleton of the system I envisioned, I thought it was important to bring the work to Mexico and try to interest librarians in that country. I could tell later from my web server logs that I had achieved some success.



Additional Sources

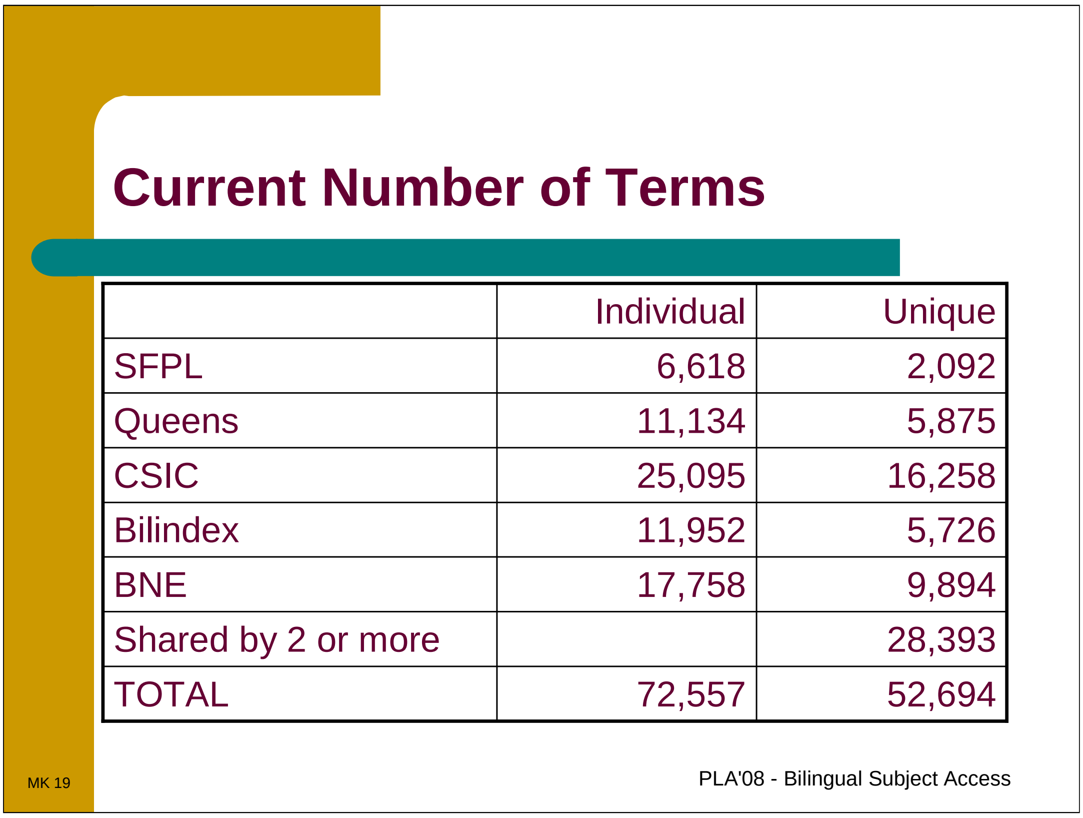
- Consejo Superior de Investigaciones Científicas (Spain)—from 2001 CD-ROM
- Biblioteca Nacional de España—harvested from online authority file, March 2007
- Bilindex 1984—from scanned index
- LCSH (for validation)—harvested by Simon Spero, Fall 2006
- Miami-Dade bib records (not yet processed)

MK 18

PLA'08 - Bilingual Subject Access

In the following months I investigated other sources of data. I had already gotten copies of bib records from Miami-Dade which I intended to process the same way as I had the San Francisco records, but I was distracted by other sources that would require less work and yield more data. I bought the 2001 CD-ROM published by the CSIC in

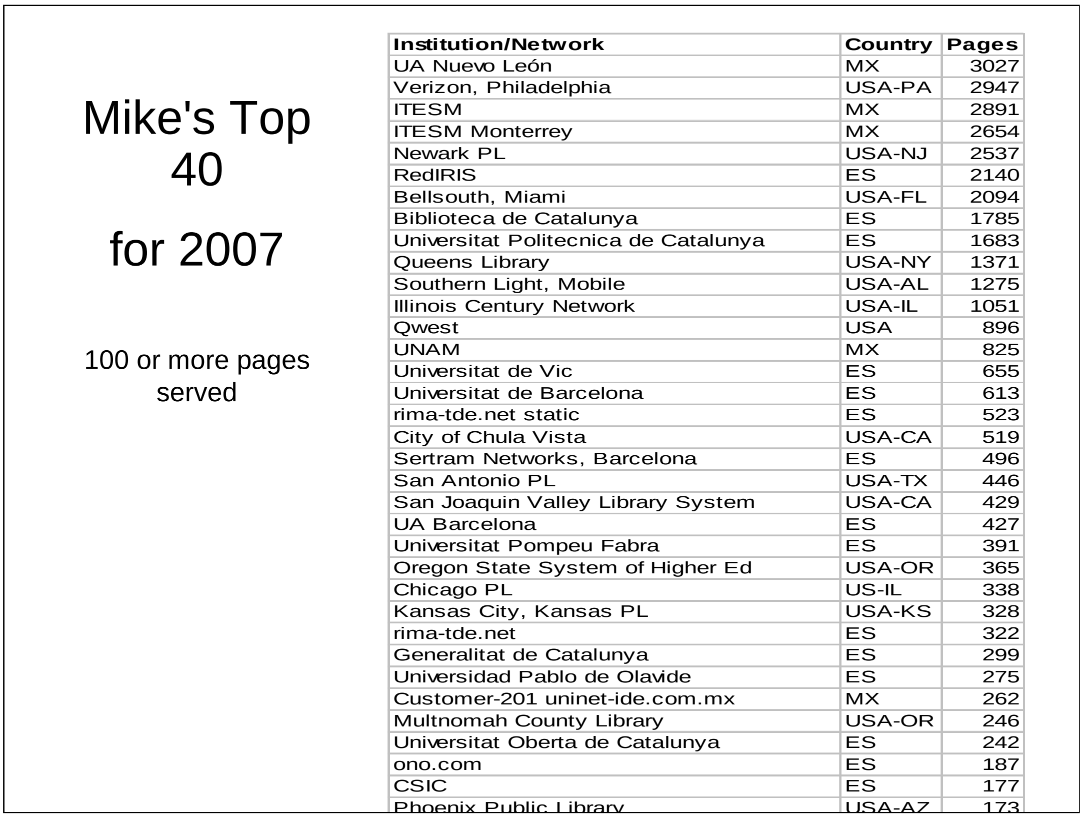
Spain and got permission to load the data. Then Simon Spero made available his downloaded LC authority records which I needed to validate LC terms from other sources, many of which were out of date. I had already been in touch with the National Library of Spain about using their data and had gotten an agreement in principle but failed to work out a mechanism for actually getting a copy of the data. I had already considered a solution like Simon's for the LC records—writing a script to download them—and after seeing his success, I decided to go ahead with it. It went quite smoothly and a couple of weekend nights netted me some 25,000 records. I also decided to scan the English to Spanish index of the original 1984 Bilindex, a fair amount of work.



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

By last summer I had accumulated over 50,000 records and had partially validated them against the LC subject authority file. The data leaves a lot to be desired. Much of it is outdated. There are still a few errors from my original processing and errors in all the sources I used, but I've been able to correct a huge number of problems. In spite of the shortcomings, traffic on the database has increased fairly steadily from the United States, Mexico, and Spain.



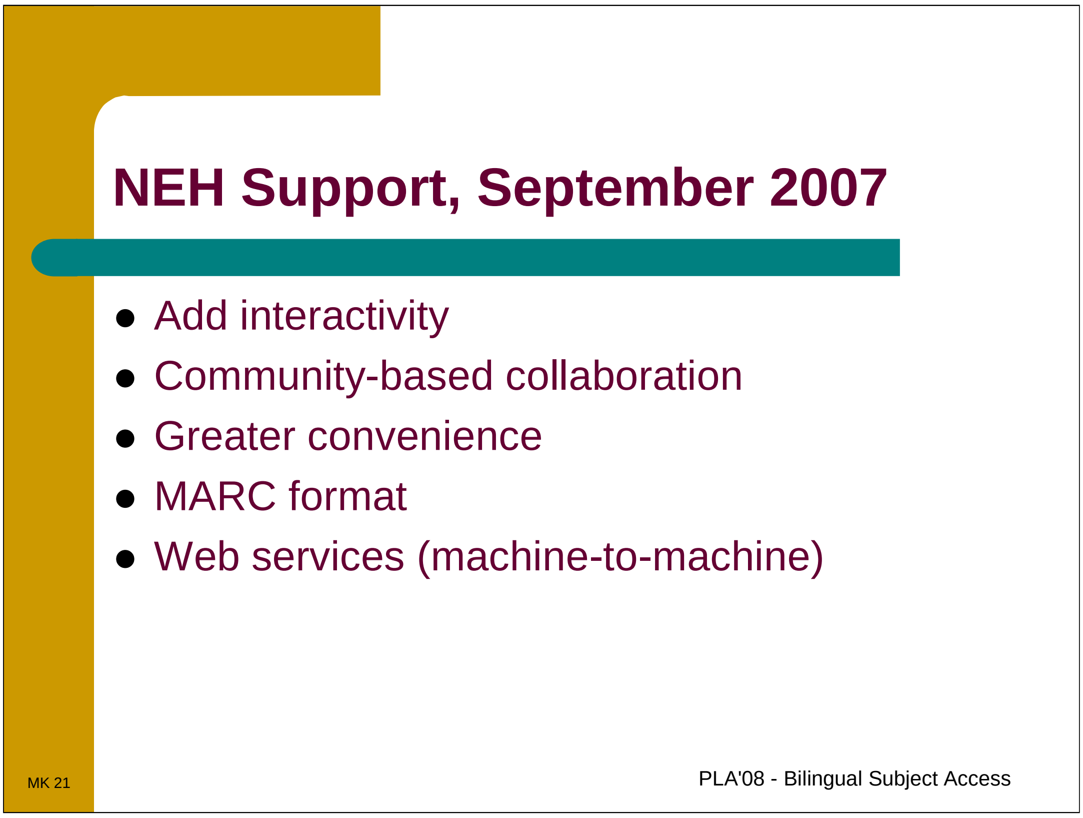
Mike's Top 40

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

Here's a list of my top 40 users for 2007. In some cases these aren't individual users, but groups of users using the same network or service provider, but it gives me a rough idea who's using the database.



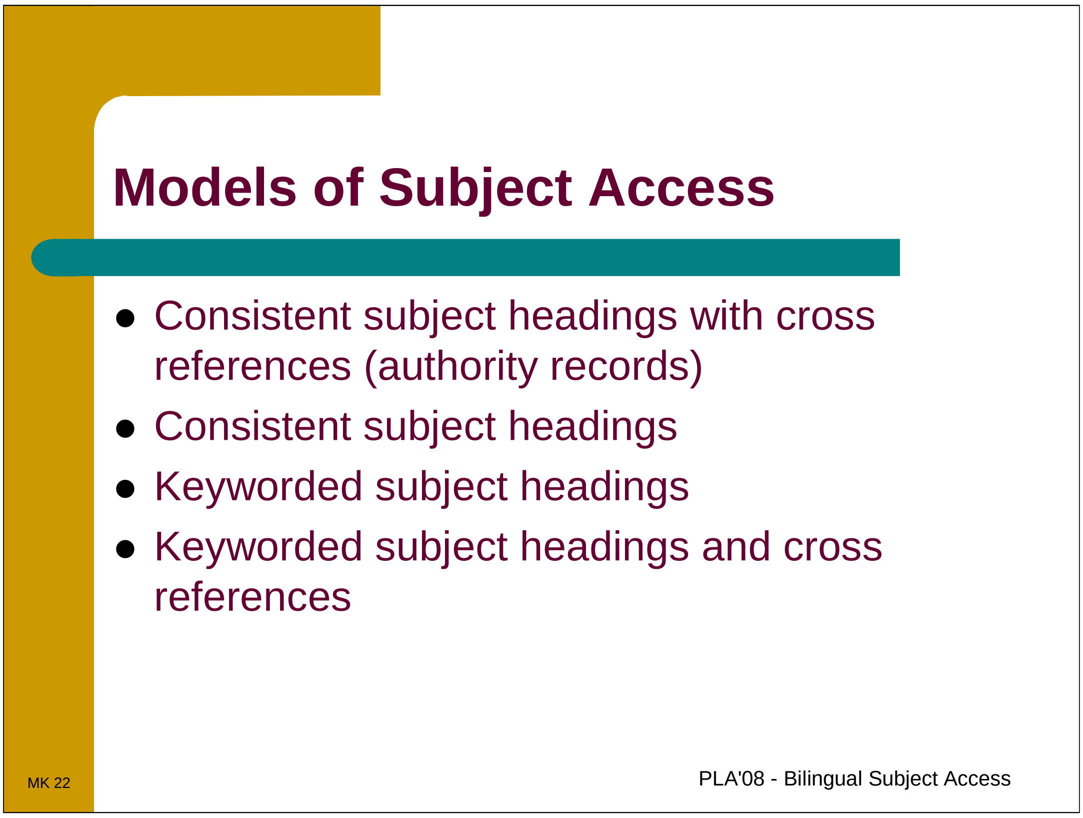
NEH Support, September 2007

- Add interactivity
- Community-based collaboration
- Greater convenience
- MARC format
- Web services (machine-to-machine)

MK 21

PLA'08 - Bilingual Subject Access

In September of last year I received a grant from the National Endowment for the Humanities to continue work on the database. This was a huge milestone for me, because until then I had to fit the work into the limited number of hours of research time I'm allotted in my day job, or else spend evenings and weekends on it. Now I can pay a student programmer to do the bulk of the work. Here you can see our current goals. The basic idea is to add interactivity and functionality, enabling catalogers, wherever they are, to contribute to the database as well as take from it. Fundamental to everything is putting all the data into MARC format and translating all characters to Unicode. This has taken a lot more time than I expected. I'd hoped to go public with a new version of the database in time for this conference, but we're not quite there. I can show you a test system. In the coming months I also plan to develop some web services for system-to-system communication and hope to offer batch services for libraries to add Spanish subject headings to their records automatically.



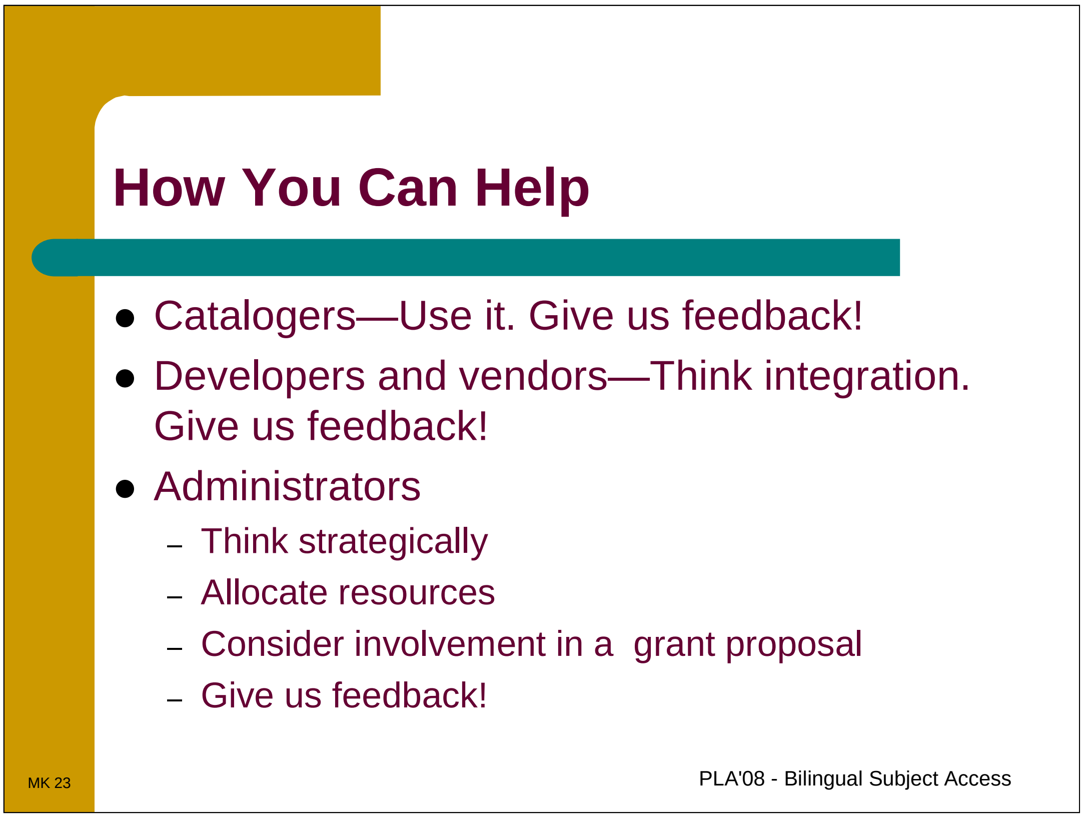
Models of Subject Access

- Consistent subject headings with cross references (authority records)
- Consistent subject headings
- Keyworded subject headings
- Keyworded subject headings and cross references

MK 22

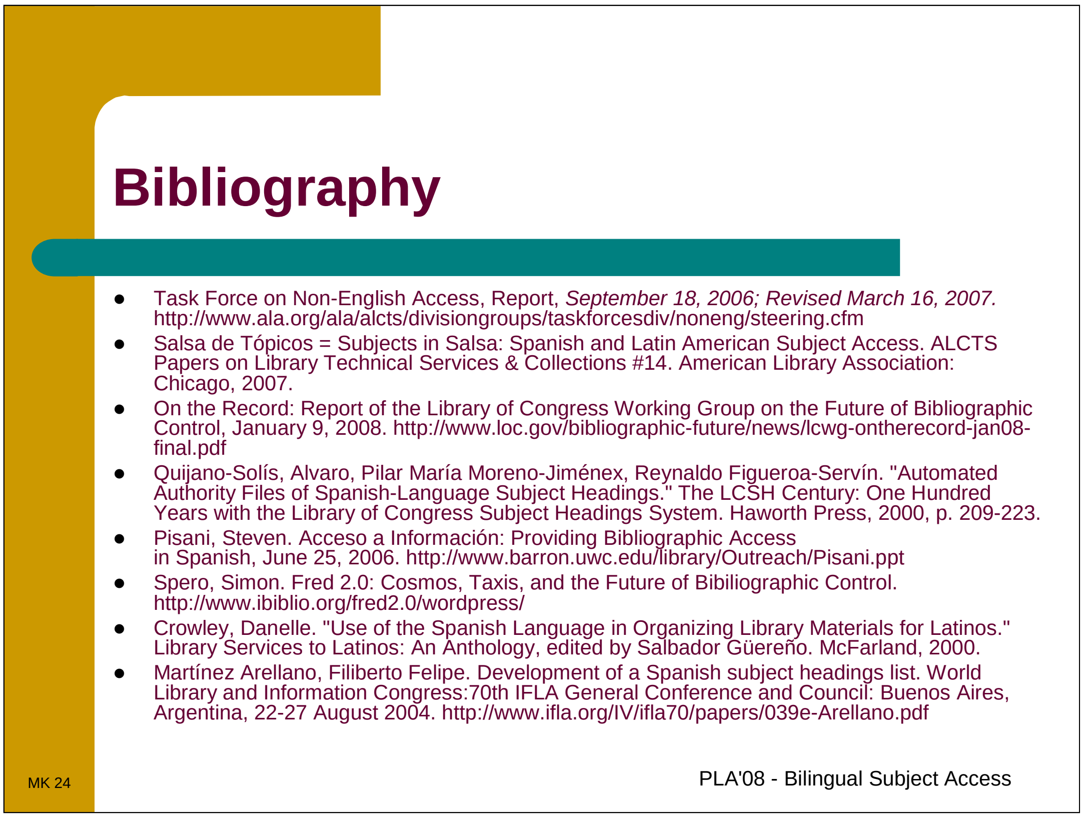
PLA'08 - Bilingual Subject Access

When it comes to putting Spanish subject headings into the catalog, I tend to think of four basic approaches, and I've listed them in descending order of sophistication. First, there's the traditional method—apply a consistent set of headings and create authority records to provide cross-references. This is the direction Queens is moving. They've built a consistent set of headings, and I'm working with them to create authority records with cross-references taken from other sources. Not too many libraries have the resources to take this route. The next step down is to apply a consistent set of subject headings, but skip the authority records. I believe this is the situation San Francisco has maintained for some years. Another approach is to take whatever headings are available, typically from OCLC or a book vendor, without worrying too much about whether they're consistent or not. These terms can be indexed specifically as subject headings, but since without consistency they probably are more useful for keyword searching. The last approach I dreamed up. Maybe someone else has thought of it, too, and maybe there's a library out there that's doing it, but it's a little crazy, so maybe not. Besides putting a subject heading into a record, why not add cross references, too, in separate fields that are keyword searchable? Disk space is the cheapest computing resource there is, so why not fatten up our records with this additional subject content? To my mind, this is a cheap and easy way to enrich bibliographic records.



How You Can Help

- Catalogers—Use it. Give us feedback!
- Developers and vendors—Think integration. Give us feedback!
- Administrators
  - Think strategically
  - Allocate resources
  - Consider involvement in a grant proposal
  - Give us feedback!



Bibliography

- Task Force on Non-English Access, Report, *September 18, 2006; Revised March 16, 2007.* http://www.ala.org/ala/alcts/divisiongroups/taskforcesdiv/noneng/steering.cfm
- Salsa de Tópicos = Subjects in Salsa: Spanish and Latin American Subject Access. ALCTS Papers on Library Technical Services & Collections #14. American Library Association: Chicago, 2007.
- On the Record: Report of the Library of Congress Working Group on the Future of Bibliographic Control, January 9, 2008. http://www.loc.gov/bibliographic-future/news/lcwg-ontherecord-jan08-final.pdf
- 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.
- Pisani, Steven. Acceso a Información: Providing Bibliographic Access in Spanish, June 25, 2006. http://www.barron.uwc.edu/library/Outreach/Pisani.ppt
- Spero, Simon. Fred 2.0: Cosmos, Taxis, and the Future of Bibiliographic Control. http://www.ibiblio.org/fred2.0/wordpress/
- Crowley, Danelle. "Use of the Spanish Language in Organizing Library Materials for Latinos." Library Services to Latinos: An Anthology, edited by Salbador Güereño. McFarland, 2000.
- Martínez Arellano, Filiberto Felipe. Development of a Spanish subject headings list. World Library and Information Congress:70th IFLA General Conference and Council: Buenos Aires, Argentina, 22-27 August 2004. http://www.ifla.org/IV/ifla70/papers/039e-Arellano.pdf