LCSH-es.org

Una Herramienta Web de Materias en Español

II Encuentro Internacional de Catalogación
12 de septiembre 2006
México, D.F.

Michael Kreyche
Kent State University

Tengo el placer y el honor de estar aquí para compartir la historia de un proyecto que me ha ocupado mucho durante el último año. Muchas veces me dormí contemplando algún detalle y amanecí a continuar pensando, si no fuera que ya me había despertado en plena noche con una nueva idea o preocupación.

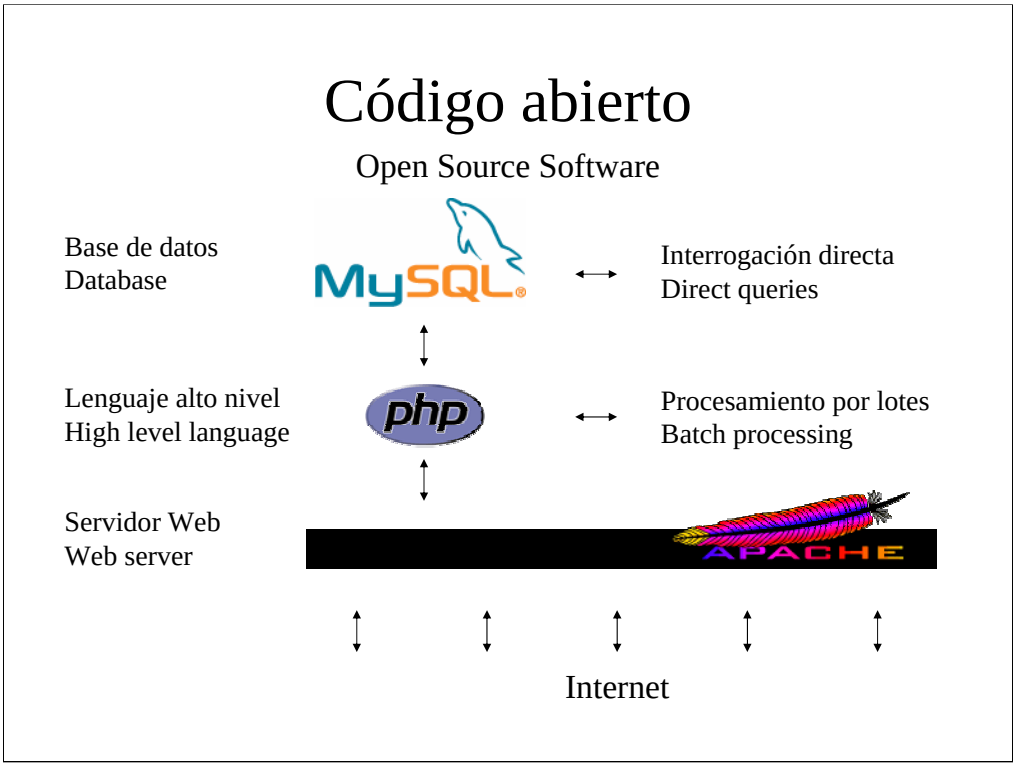
Las ideas que nos agarran con tanta fuerza suelen plantearse mediante la inspiración de una persona a quien estimamos y quien nos tiene confianza. En este caso, la inspiración fue una profesora de catalogación. Cuando cursé la materia hace unos veinticinco años, me impresionó mucho la Dra. Ann Allan y siguió siendo una inspiración aún después de su jubilación.

Dra. Ann Allan

- Profesora de catalogación
- Voluntariado en Costa Rica
- Apoyo al taller LCSH en 2002 (Washington)

Ella viajó varias veces a Costa Rica donde prestó ayuda voluntaria en la biblioteca de CATIE, Centro Agronómico Tropical de Investigación y Enseñanza. Por varios años se ha esforzado a generar apoyo moral y financiero para una traducción de los Encabezamientos de Materia de la Biblioteca del Congreso (LCSH) al español.

Hace poco más de un año pidió mi ayuda. Me entregó una lista de sus contactos y yo me comprometí a hacer algo.



Como mi pericia no está en la catalogación sino en sistemas, me puse a pensar en como apoyar tal proyecto con recursos informáticos. Dentro de pocas semanas procuré configurar un sistema utilizando tecnología Web con la base de datos MySQL, el servidor Apache, y el lenguaje de programación PHP para crear páginas dinámicas. Todos son productos de código abierto de alta calidad.

San Francisco Public Library

Vivian Pisano

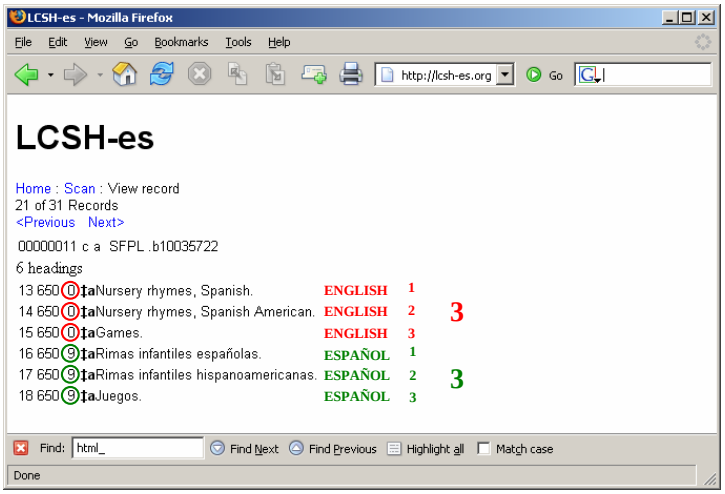
- Participó en la produccion del Bilindex
- Bibliotecas Para La Gente (REFORMA)
  (mantenía Lista Bilíngüe de materias en sitioWeb)
- 15,000 registros bibliográficos con materias en español

Mi primera comunicación con los contactos de Dra. Allan fue con Vivian Pisano de la Biblioteca Pública de San Francisco, California, quien había participado en la producción del Bilindex. Ella estuvo muy de acuerdo con la idea de una traducción actualizada de LCSH.

Propuse la idea de analizar registros bibliográficos con materias en los dos idiomas para derivar un diccionario bilingüe de términos. En efecto, recrear una parte del Bilindex mediante los registros bibliográficos producidos con él. Ella tuvo la bondad de compartir datos y en poco tiempo recibí unos 15,000 registros.

Agrupados por Idioma

Arranged by Language



Cargué los encabezamientos de materia en mi base de datos y vi que los registros se habían elaborado con bastante cuidado y consistencia. La mayoría exhibe ciertas características.

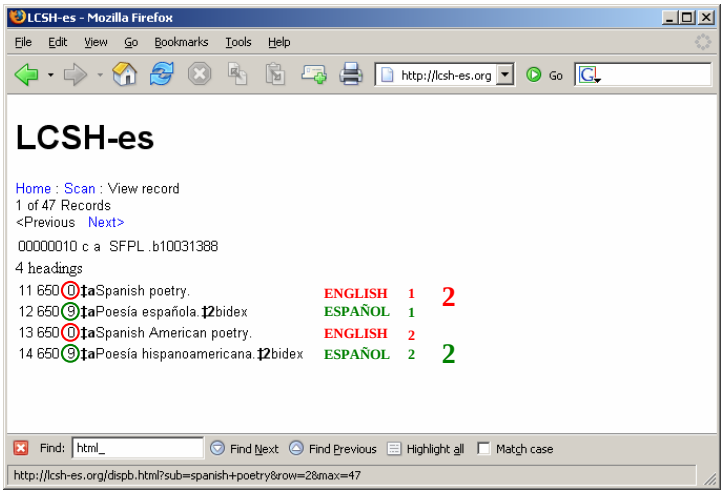
En primer lugar, se puede distinguir el idioma de cada asiento por el segundo indicador MARC. Esto es de suma importancia para un análisis computacional.

~Por otra parte, cada registro generalmente tiene el mismo número de encabezamientos en español que en inglés.

~Y tercero, se mantiene el mismo orden en los dos idiomas, según dos patrones: el grupo de los en inglés seguidos por el grupo de los en español...

Agrupados por Encabezamiento

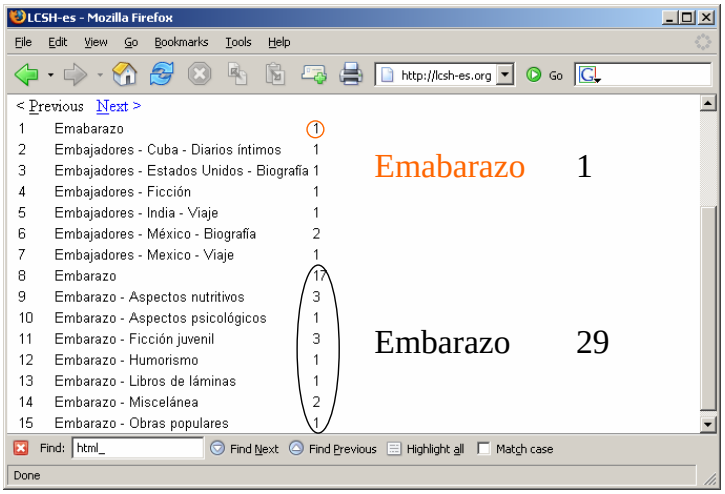
Arranged by Heading



..., o, cada encabezamiento en inglés seguido por el correspondiente en español.

Entre más datos, más confianza

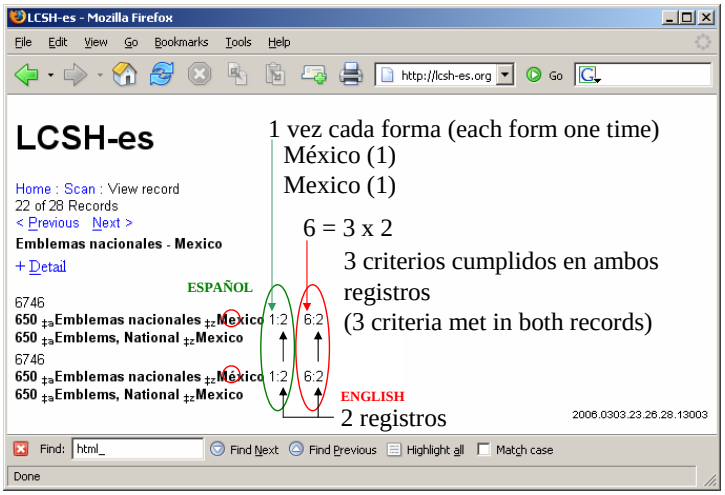
More data, greater confidence



La siguiente tarea fue indizar los asientos para poder juntar todas las ocurrencias de cada materia y analizar los contextos en que se encontraban. Entre más ocurrencias de un dado encabezamiento, más información habría a analizar y más seguridad en las conclusiones.

Sistema de calificación

Rating system



Luego comenzó otra tarea más larga—el análisis de los asientos en español con el propósito de inferir los correspondientes en inglés. Hice una interfaz dinámica para hacer pruebas y visualizar paso a paso los resultados del algoritmo que iba desarrollando. Aquí se ve un ejemplo con el resumen de la calificación de las correspondencias.

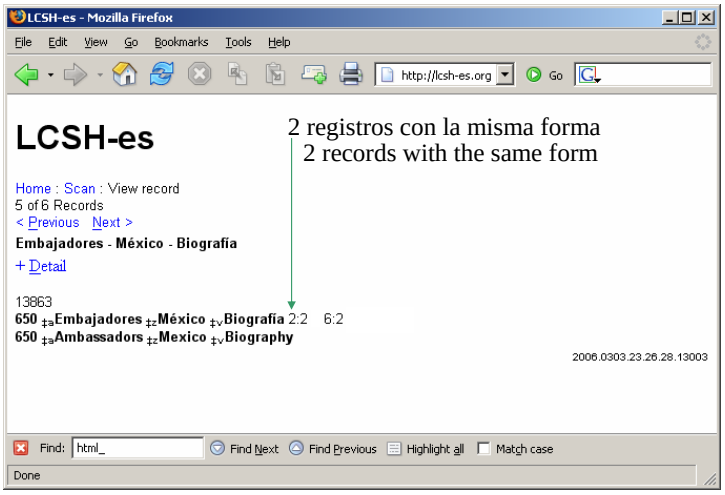
Hay dos calificaciones, una para el español y otra para el inglés.

~Las dos se representan por una proporción a base del número de registros, en este caso dos.

~La proporción 6:2 tiene que ver con el inglés y es la parte más importante. El dos refiere a los dos registros y el seis representa tres criterios que se cumplen en ambos (o sea, tres por dos son seis).

~La proporción 1:2 quiere decir que de los dos registros una sola forma del español aparece en cada uno.

Caso sencillo



Aquí tenemos otro ejemplo más sencillo, sin errores de acentuación.

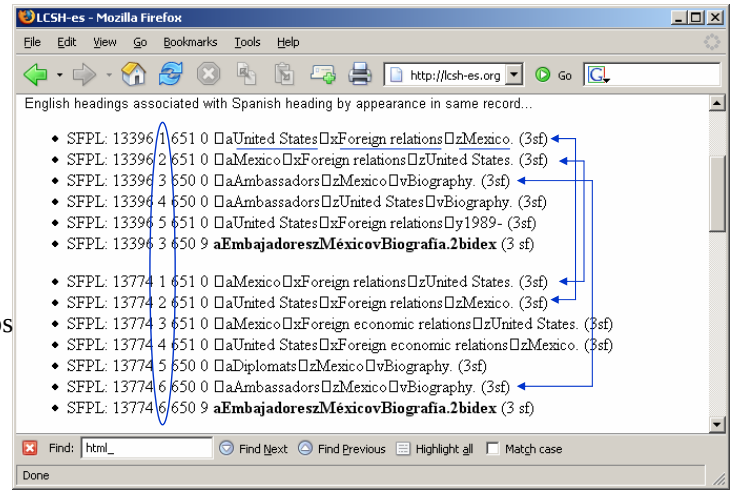
3 Criterios

3 criteria

1. Orden

2. Frecuencia

3. No. de subcampos



Después de mucha experimentación, concluí tomando en cuenta tres criterios. Dos de los criterios se relacionen con las características ya mencionadas: el respectivo orden de los asientos

~y la frecuencia de los varios asientos en inglés entre los registros donde se encuentra un dado asiento en español.

~El tercero es el número de elementos o subcampos dentro del campo.

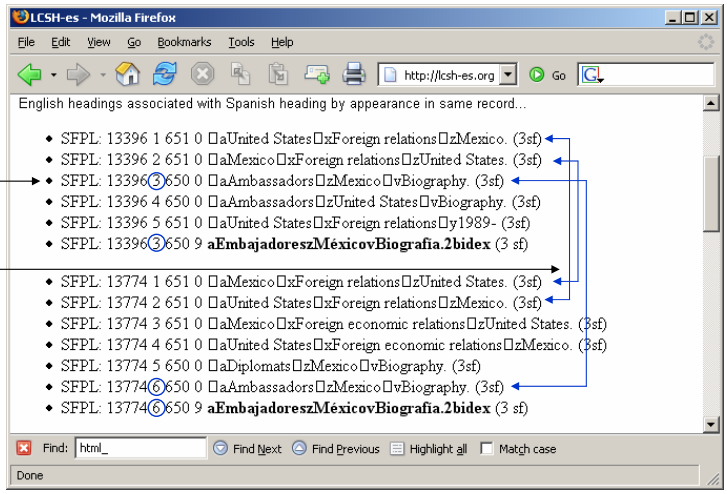
Ejemplo de Calificación

Rating Example

Orden:
1 con “2”

Frecuencia:
3 con “2”

No. de subcampos:
3 con “2”

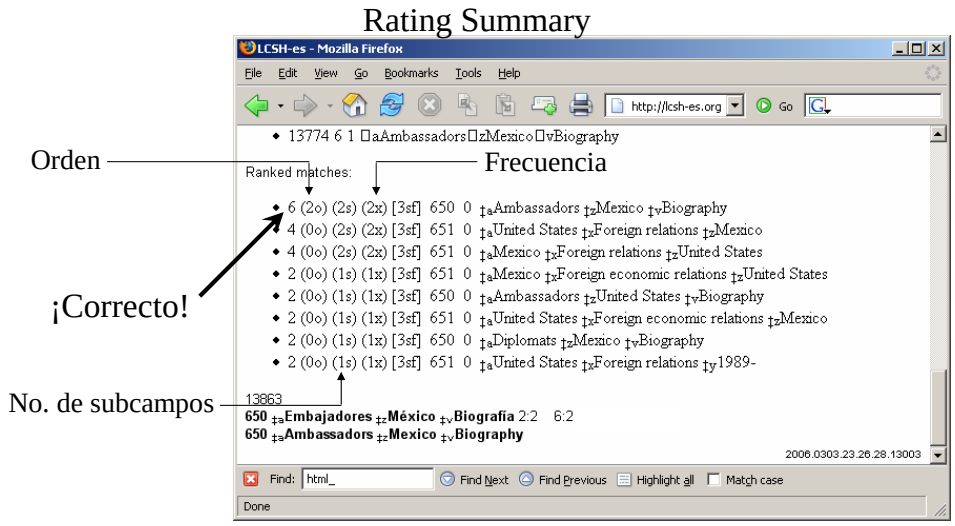


Veamos más detalles del análisis. En los dos registros “Embajadores—México—Biografía” tiene el mismo orden que el encabezamiento en español. Y por eso le asignmaos “2” en cuanto al orden.

Este y dos otros ocurren en ambos registros y por lo tanto les asignamos el valor de “2” en cuanto a la frequencia.

Todos tienen dos subdivisiones, o sea tres elementos o subcampos en total, igual que el encabezamiento en español. Algunos se merecen “1” y otros “2” en este respecto por occurrir sólo una o dos veces.

Resumen de Calificaciones



El programa produce un resumen de todas las calificaciones, ordenado según la suma total de las calificaciones.

Revisión de Correspondencias

Review of Matches

- Muestra de 400
- 25 errores (6%)
- Con 24 errores, existía también una correspondencia correcta

Cuando me sentí satisfecho con los resultados y me cansé de mejorar el algoritmo, hice una versión del programa que escribiera los resultados a la base de datos y me puse a evaluarlos. De los 15,000 registros bibliográficos se generaron 8,905 correspondencias. Imprimí una muestra de 400 registros y los examiné uno por uno. Entre estos, encontré 25 errores o un 6%. Lo más interesante es que con 24 de las 25 correspondencias equivocadas, existía también una correspondencia acertada.

Ejemplos de Pares

Dance	Música popular
Dance	Danza
Early childhood education	Educación infantil temprana
Early childhood education	Literatura infantil
Games	Juegos
Games	Mitología

Por ejemplo, encontré Dance y Música popular pero también Dance y Danza. Eso implica que para eliminar la mayoría de los errores, solo tuve que revisar los registros duplicados—

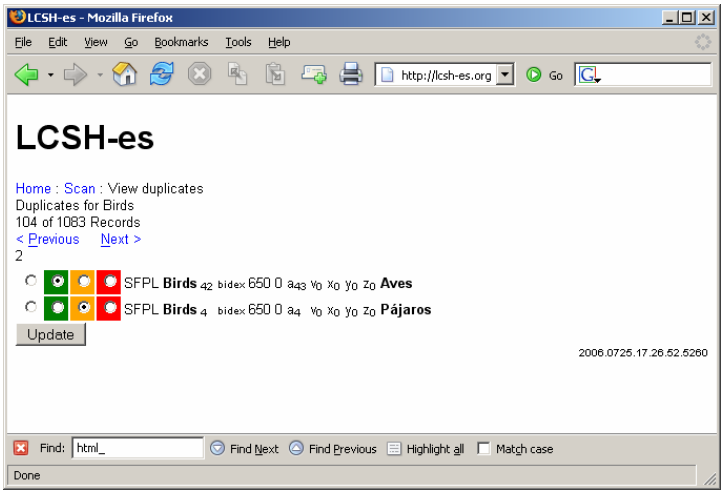
Ejemplos de Pares

Dance	Danza
Early childhood education	Educación infantil temprana
Games	Juegos

Y eliminar las correspondecias erróneas. Por eso hice una interfaz a la base de datos que identificara las correspondencias duplicadas y facilitara la eliminación de las falsas.

Resolución de Duplicados

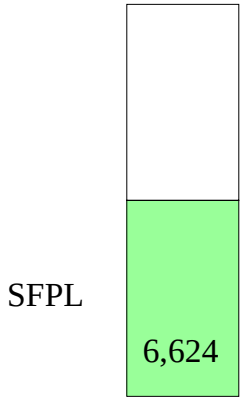
Resolving Duplicates

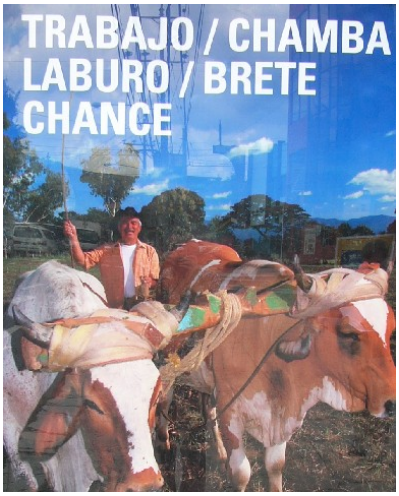


loc rem

Esta permitió indicar el estado de cada correspondencia: bueno, malo, o dudoso. En dos días revisé los 2,200 pares de correspondencias contradictorias y eliminé las evidentemente falsas. Quedan aproximadamente 200 pares que representan diferentes traducciones del mismo encabezamiento. Estos se pueden resolver más tarde, o por eliminación o por conversión en referencias.

Diccionario de Términos



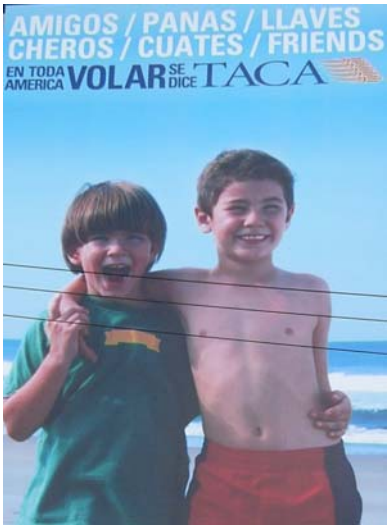


De esta depuración resultaron 6,624 términos.

Desde este punto voy a incluir varias fotos de anuncios de la aerolínea TACA. Estuve en El Salvador en julio y no pude dejar de pensar en los encabezamientos de materia porque por todos lados se veían estos anuncios.

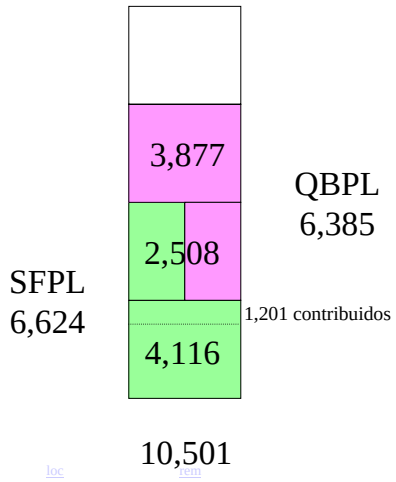
Queens Borough Public Library

- Extracción de asientos de materia en inglés
- Diccionario bilingüe
- Formato XML
- 3/4 traducido



Mientras tanto, Dra. Allan se reunió con los administradores de la Biblioteca Pública de Queens Borough (Nueva York) quienes autorizaron una colaboración. Llevaban a cabo un proyecto para agregar asientos en español a su catálogo, mediante la creación de un diccionario de términos en formato XML. Cuando este diccionario quede completo, los registros bibliográficos se modificarán con un procesamiento por lote. Como coincidíamos en el mismo concepto, la construcción de un diccionario, fue muy fácil compartir datos.

Aumento de la Base de Datos



Resultaron 10,501 registros. De estos, las dos bibliotecas tienen 2,508 en común, o sea que están de acuerdo en la traducción del inglés. Del resto, 3,877 son de Queens y representan un aumento de la base de datos. 4,116 son de San Francisco y de ellos 1,201 fueron agregado al diccionario de Queens.

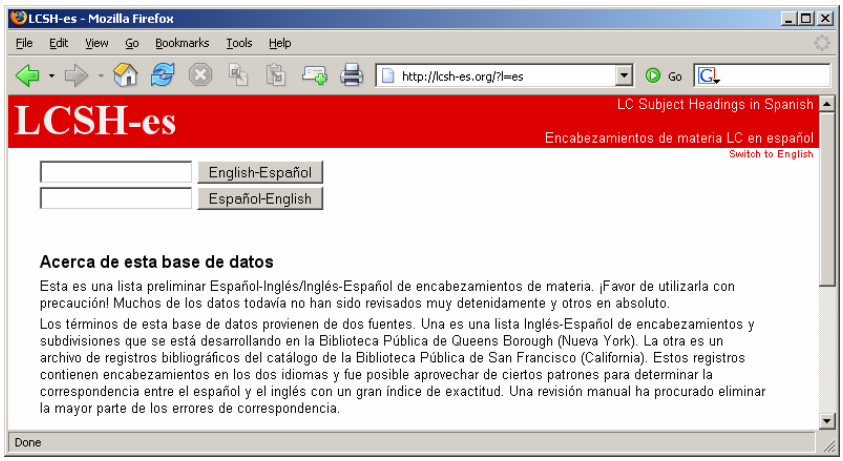
Además, la base se puede analizar para encontrar duplicados contradictorios de la misma manera en que se detectaron las falsas correspondencias obtenidos del catálogo de San Francisco.

Beneficios de Colaboración

Aumento de la base de datos Increase to the database	3,877	58% del total de San Francisco
Equivalencias Confirmadas Confirmed Matches	2,508	24% del total de ambas bibliotecas
Contribución de datos a QBPL Contribution to Queens	1,201	40% de lo que faltaba a Queens
Equivalencias falsas o alternativas False or alternative matches identified	1,083	10% del total de ambas bibliotecas

En suma, esta colaboración ha producido cuatro beneficios: aumento de registros; confirmación de equivalencias; contribución de datos al diccionario de Queens; y posibles equivalencias alternativas.

lcsh-es.org



loc rem

¿De aquí en adelante, que queremos lograr? Primero, a corto plazo, una herramienta gratuita que contribuya a satisfacer las necesidades de bibliotecas que tienen o encuentran registros con encabezamientos LC y quieren agregar los equivalentes en español. Desde mi punto de vista, faltan dos cosas para lograr esto.

Mejoramientos

Improvements

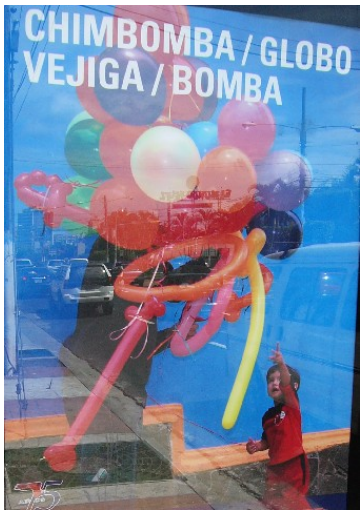
Búsquedas múltiples

650 0 ‡aNursery rhymes, Spanish.
650 0 ‡aNursery rhymes, Spanish American.
650 0 ‡aGames.

↓

650 9 ‡aRimas infantiles españolas.
650 9 ‡aRimas infantiles hispanoamericanas.
650 9 ‡aJuegos.

Multiple searches



Primero, más funcionalidad. En este momento, hay que buscar cada elemento uno por uno. Lo que tengo planeado es un cuadro donde uno pudiera pegar una serie de encabezamientos LC y con un solo clic recuperar los equivalentes, completos con subdivisiones y, opcionalmente, códigos MARC, listos para copiar y pegar al sistema del usuario.

Esto sería una búsqueda manual. No lo veo tan difícil y espero desarrollar una primera versión en los últimos meses de este año. Más adelante me gustaría desarrollar un servicio Web para poder automatizar esta función.

Esto depende de la voluntad de otra persona o empresa a desarrollar el software cliente por el lado del sistema local. Otra posibilidad es facilitar conversiones por lote con el software que se está desarrollando para Queens, lo cual será distribuido gratis.

Más Datos

More Data

•Miami-Dade Public Library System (40,000 registros bibliográficos)

•Escaneo del índice del Bilindex?

La segunda falta es de datos. Claro que si no hay una suficiencia de datos, las búsquedas no pueden tener mucho éxito. Por eso sigo buscando contribuciones de datos.

Hace pocas semanas recibí un archivo de registros bibliográficos del Sistema Bibliotecario Público de Miami-Dade (Florida). A primera vista promete aumentar la base de datos otro treinta por ciento, si puedo superar algunos problemas con los datos. Pienso desarrollar otro algoritmo mejorado para encontrar las correspondencias entre el inglés y el español, aprovechando los datos ya organizados en la base.

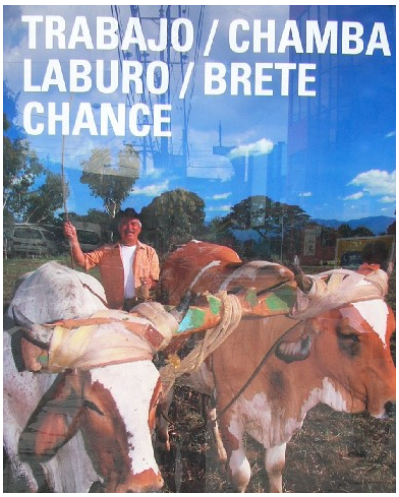
Estoy pensando también en escanear los índices del Bilindex original.

Lo ideal sería obtener registros de autoridades que contengan el término autorizado en los dos idiomas, o por lo menos un punto de enlace como el número de control del correspondiente registro LC.

Más Datos—Autoridades

More Data--Authorities

- Westchester Library System (Nueva York)
- Biblioteca Nacional de España
- LEMB?
- México?
- Otros?



El Sistema Bibliotecario de Westchester (Nueva York) está construyendo tal archivo de registros de autoridades en español con la intención de compartir los datos cuando se termine el proyecto.

También he investigado las dos fuentes comerciales de registros de autoridades en español de que estoy enterado, el CD de ProQuest con el archivo de la Biblioteca Nacional de España (BNE) y el LEMB Digital.

Desgraciadamente, las licencias de los vendedores prohiben que se compartan los datos fuera de la institución. Ya que los registros de autoridades de la BNE se pueden descargar de la Web, hay otra opción para obtenerlos, y he iniciado comunicación con la BNE para buscar otra manera más adecuada para los miles de registros que quisiera cargar. Así mismo pienso comunicarme con la Biblioteca Luis Ángel Arango para investigar acerca de sus registros.

Aparte de los casos ya mencionados, sé bien que aquí en México y en otros países hay varios esfuerzos a crear y mantener registros de autoridades. Uno de los propósitos de venir aquí es conocer más sobre ellos. No tengo duda de que estos recursos serían de enorme utilidad para bibliotecas externas—si se compartieran—y pido colaboración en la forma de contribuciones de datos.

Principios y Políticas

Principles and Policies

"Es necesario, antes de iniciar cualquier proyecto de autoridades, generar los principios y las políticas de formación de las mismas. Mismos que deben responder a las necesidades de las comunidades a las que se les da servicio."

Fernando Alvarez Ortega
Primera Reunión Nacional sobre Control de Autoridades
Colegio de México
17 de marzo de 2001

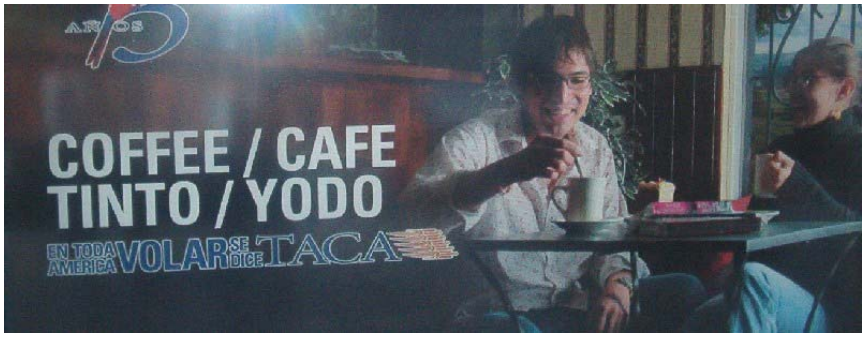
(también alumno de Ann Allan)

Además de la meta inmediata de construir una herramienta útil de acceso libre, estamos dispuestos a colaborar en la producción de una adaptación definitiva de LCSH en español y estamos buscando fuentes de financiamiento aunque sea para un proyecto de alcance limitado, como piloto.

Las líneas generales de una adaptación de LCSH para Latinoamérica han sido expuesto por otros, entre ellos en el año 2004 por Dr. Martínez Arellano y Ageo García. Cabe señalar también los comentarios de Fernando Álvarez Ortega en una reunión aquí en México en 2001 sobre el control de autoridades, no sólo por el valor de sus observaciones sino también porque él era estudiante de la Profesora Allan. Estoy completamente de acuerdo con él en que, "Primero. Es necesario, antes de iniciar cualquier proyecto de autoridades, generar los principios y las políticas de formación de las mismas. Mismos que deben responder a las necesidades de las comunidades a las que se les da servicio."

Por eso debe ser una prioridad organizar una conferencia con este solo propósito: exponer, discutir y debatir desde varios puntos de vista, los principios en que basar los encabezamientos de materia en español.

Temas de Discusión



Para terminar quisiera ofrecer algunas observaciones generales.

No tenemos tiempo suficiente para explorar a fondo el concepto, todavía impreciso, de lo que es la Web 2.0, pero quisiera exponer algunas de los puntos que se le atribuyen, adaptados a la construcción de un tesauro de materias en español vinculado a LCSH. No están bien desarrolladas estas ideas y las ofrezco como temas de discusión.

Tesauro en la Web 2.0

- La Web es La Plataforma
- El Tesauro es Servicio, Sencillo y Ligero
- El Tesauro es Adaptable a Prácticas Locales o Nacionales
- El Tesauro Tiene Arquitectura de Participación
- El Tesauro Está en Beta Perpetuo

•La Web es la plataforma: La manifestación primaria de las autoridades está en la Web y no integrado con un catálogo bibliográfico. Para las bibliotecas que todavía no tienen acceso Internet, se pueden derivar otros formatos.

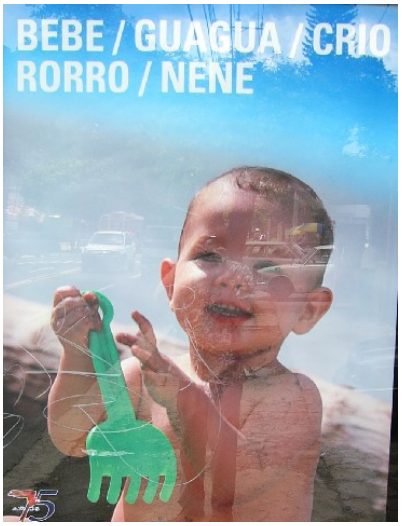
•El tesauro es servicio: El sistema mantiene la sencillez y rapidez de una herramienta especializada. Se puede consultar de manera manual o puede suministrar datos a otros sistemas por protocolos livianos y ligeros.

•El tesauro es adaptable: La estructura sigue las normas internacionales con extensiones que acomodan prácticas locales o nacionales. Cada biblioteca puede configurarlo según sus necesidades.

•El tesauro tiene una arquitectura de participación: Los usuarios (bibliotecas) contribuyen valor agregado. Participan en la elaboración de los datos con comentarios específicos y extractos de datos institucionales. Entre más datos y más participación, mejor el servicio.

•El tesauro está en Beta perpetuo: Tenemos un recurso útil desde el inicio en un estado de mejoramiento contínuo. Tal vez no contamos con versiones o ediciones numeradas.

Nueva Generación de Tesauro



Estos son los elementos básicos de mi visión y espero seguir incorporándolos en LCSH-es. Junto con mis colaboradores de las otras bibliotecas, ofrezco esta base de datos, http://lcsh-es.org, a la comunidad bibliotecaria y invito su exploración y participación para que probemos estos conceptos. Si tiene su institución datos que puedan expandir su alcance, pido que piense cómo compartirlos y bajo qué condiciones.

Tenemos una historia en la cual abundan ejemplos de listas de encabezamientos de materia. Cada una a su vez valió para la elaboración de la siguiente generación o versión de herramienta temática. Ya es hora de encontrar una manera de unir los esfuerzos dispersos y dar a luz a una nueva generación de tesauro.