# Seamless Searching of Numeric and Textual Resources
# Final Report on Institute of Museum and Library Services National Leadership Grant No. 178

Principal Investigators: Michael K. Buckland, Fredric C. Gey and Ray R. Larson
School of Information Management and Systems
University of California, Berkeley, CA 94720-4600

December 20, 2002

## 1 Summary

This three year project demonstrated improved access to written material and numerical data on the same topic when searching two very different kinds of databases (books, articles, and their bibliographic records) and numerical data (socio-economic databases). Also, search support was developed for transverse searching whereby data found in a text database can be used to find related data in a numeric database and vice versa.

## 2 Introduction

A hope for new technology in libraries has been to support seamless searching across an increasing range of resources on a growing digital landscape. The reality is that network-accessible resources, like the contents of a well-stocked reference library, are quite heterogeneous, especially in the variety of indexing, classification, categorization, and other forms of "metadata."

The intent of this project is to demonstrate improved access to written material and numerical data on the same topic when searching two quite different kinds of database: text databases (books, articles, and their bibliographic records) and numerical data (socio-economic databases).

The problem is that there has, until now, been no easy path to integrate numeric databases with bibliographic and textual databases which might contain knowledge about cause and effect. The vocabulary which classifies the numeric data may be quite different from the subject headings used for books, magazine articles, and newspaper stories about the same topic of interest. Also there needs to be an environment of search support that facilitates such transverse searching, establishing connections, transferring data and invoking appropriate utilities in a helpful way.

This project addresses both problems through the development and demonstration of a library gateway providing search support for both text and socio-economic numeric databases. The gateway will help users conduct searches in each type of database by accepting a query in the library users' own terms and then suggesting the specialized categorization terms to search for in the information resource (database). The intent is that if you found something interesting in a socio-economic database, the gateway will help you to find documents on the same topic in a text database, and vice versa. Selection of the best search terms in the target databases is supported by the use of "Entry Vocabulary Indexes," which resemble Melvil Dewey's "Relative Index," but are created using statistical association techniques.

## 3 Data sets

We obtained a copy of the MARC files from MELVYL, the University of California online catalog. We selected 4,246,510 records containing at least one subject heading (6xx field). From this set of records, we extracted the fields containing titles and subtitles (245, subfields a and b), summaries describing the scope and general content of the material (520, subfield a), and main subject headings (6xx, subfield a). Two sample records are shown below.

```
Sample record 1:
<DOC>
<001>73180254 //r86</001>
<245><a>A study of operant conditioning under delayed reinforcement
in early infancy</a></245>
<650><a>Infant psychology.</a></650>
<650><a>Operant conditioning.</a></650>
</DOC>

Sample record 2:
<DOC>
<001>73180255 </001>
<245><a>Reptilian disease: recognition and treatment,</a></245>
<650><a>Reptiles</a><x>Diseases.</x></650>
</DOC>
```

## 4 Association Measure

We used statistical association methods to provide a path from the words in the query to the topical classification (classification numbers, category codes, subject headings, etc.) in the data set. The final stage to creation of an Entry Vocabulary Index is to develop a maximum likelihood ratio weighting associated with each term (word or phrase) and each metadata value. One constructs a two-way contingency table for each pair of word/phrase terms $t$ and classification codes $C$ as shown in table 1. where $a$ is the number of documents with titles containing the word or phrase and classified by the classification code; $b$ is the number of documents with titles containing the word or phrase but not classified by the classification code; $c$ is the number of documents with titles not containing the word or phrase but classified by the classification code; and $d$ is the number of documents with titles not containing the word or phrase and not classified by the classification code.

|  | $C$ | $\neg C$ |
|---|---|---|
| $t$ | a | b |
| $\neg t$ | c | d |

Table 1: Contingency table from words/phrases to classification

The association score between a word/phrase $t$ and an classification $C$ is computed following Dunning [3]

$$W(C,t) = 2[logL(p_1,a,a+b) + logL(p_2,c,c+d) - \tag{1}$$

$$logL(p,a,a+b) - logL(p,c,c+d)] \tag{2}$$

where

$$logL(p,n,k) = klog(p) + (n-k)log(1-p) \tag{3}$$

and

$$p_1 = \frac{a}{a+b}, \qquad p_2 = \frac{c}{c+d}, \qquad and \qquad p = \frac{a+c}{a+b+c+d}.$$

Metadata can be Library of Congress Subject Headings (LCSH), Library of Congress Classification Numbers, U.S. Patent Classification Numbers, and so on. One can create not only associative dictionaries that will map words in natural languages into metadata terms, but also, in reverse, associative dictionaries that will return words in natural language that are closely associated with a metadata term. If there are training records containing texts in two languages, then one could create associative dictionaries that will, in response to words as query words in one language, return words in the other language that are closely associated with the query words.

In addition to the maximum likelihood ratio-based association measure, there are a number of other association measures, such as the Chi-square statistic, mutual information measure, and so on, that one can employ in creating association dictionaries.

## 5 Entry Vocabulary Indexes

Entry vocabulary indexes (EVIs) are associative dictionaries that map vocabularies from one language to another. We have created an entry vocabulary index that maps words in natural language to the Library of Congress Subject Headings (LCSH) that are most closely associated with the query words, and an entry vocabulary index that maps a LCSH to words that are most closely associated with each main heading in the LCSH.

### 5.1 Pre-processing

The training set used to create this word-to-LCSH entry vocabulary index contains the bibliographic records with at least one assigned Library of Congress Subject Heading (i.e., at least one 6xx field). The words are extracted from the subfields *a* and *b* in the title field *245*, and from the subfield *a* in field *520*. The texts in subfields *a* and *b* of field 245, and in subfield *a* of field 520 are tokenized; the stopwords are removed; and the remaining words are normalized. A token here can contain only letters and digits. All tokens are then changed to lower case. The stoplist has about 600 words that are considered not content-bearing words, such as pronouns, prepositions, coordinators, determiners, and the like. We refer to the words that are not stopwords as content words. The content words are normalized using a table derived from an English morphological analyzer [2]. The table maps plural nouns into singular ones; verbs into the base form (the infinitive form); and comparative and superlative adjectives to the positive form. For example, the plural noun *printers* is reduced to *printer*, and *children* to *child*; the comparative adjective *longer* and the superlative adjective *longest* are reduced to *long*; and *printing*, *printed* and *prints* are all reduced to the same base form *print*. When a word belonging to more than one part-of-speech category can be reduced to more than one form, it is changed to the first form listed in the morphological analyzer table. As an example, the word *saw*, which can be a noun or the verb *see* in the past tense, is not reduced to *see*. The subject headings are extracted from the subfield *a* of fields 600, 610, 611, 630, 650, and 651. The subject headings are changed to lower case. Each subject heading in the subfield $a$ of a 6xx field is treated as one unit in creating the word to LCSH entry vocabulary index. The texts in MARC records are encoded in MARC-21 encoding. The characters with diacritic marks are encoded in two bytes, one byte for the base character followed by another for the diacritic mark. In pre-processing the data, we did not remove the diacritic marks.

From this training set of MARC records we created a word to LCSH entry vocabulary index using the statistical association measure described in section 4.

## 5.2 Word to LCSH Entry Vocabulary Index

A word to LCSH entry vocabulary index maps a word to LCSHs that are most strongly associated with the query word. As an example, the following table presents the top-ranked ten LCSHs (only the subfield $a$ in lower case) that are most strongly associated with the query word *alcoholism*.

| Rank | LCSH (subfield *a*) | Weight |
|---|---|---|
| 1 | alcoholism | 7470.46 |
| 2 | alcoholic | 1745.23 |
| 3 | alcohol | 709.26 |
| 4 | alcoholism and employment | 318.26 |
| 5 | drug abuse | 257.75 |
| 6 | alcohol, ethyl | 235.13 |
| 7 | drinking of alcoholic beverages | 151.46 |
| 8 | substance abuse | 146.04 |
| 9 | children of alcoholics | 129.53 |
| 10 | social work with alcoholics | 73.91 |

An entry vocabulary index takes as query a text fragment which can be a single word, a phrase, a set of keywords, an incomplete (or complete) sentence, a book title, and so on. The query is processed in the same way as the texts extracted from MARC fields 245 and 520 when the dictionary was created. The query is first tokenized, stopwords removed, and then the content words normalized. The set of normalized words are submitted to the entry vocabulary index. For each word, a list of ranked LCSHs with association weights, like the one shown in the table above, is generated. The ranked lists of LCSHs for all the content query words are combined to create a unified ranked list of LCSHs for the whole query. No distinction is made between languages, so foreign words found in the titles of books in foreign languages are included. For example, a query using the word *Wirtschaftspolitik*, which means *economic policy* in German, leads appropriately to the following subject headings:

| Rank | LCSH (subfield *a*) | Weight |
|---|---|---|
| 1 | economic policy | 756.90 |
| 2 | germany (west) | 645.02 |
| 3 | switzerland | 97.70 |
| 4 | regional planning | 96.39 |
| 5 | economics | 92.14 |

As an example of a query having more than one word, the top-ranked ten LCSHs (only subfield $a$ in lower case) that are most strongly associated with the query *peanut butter* are listed in the table below.

| Rank | LCSH | Weight |
|---|---|---|
| 1 | peanut | 1343.90 |
| 2 | cookery (peanut butter) | 429.61 |
| 3 | cookery (peanuts) | 423.47 |
| 4 | peanut industry | 359.57 |
| 5 | peanut butter | 316.23 |
| 6 | butter | 309.36 |
| 7 | schulz, charles m | 277.30 |
| 8 | cookery | 197.08 |
| 9 | peanut products | 170.05 |
| 10 | peanut craft | 123.15 |

The ranked list of LCSHs for the query *peanut butter* was produced by combining the ranked list of LCSHs for the query word *peanut* and that for the query word *butter*. In combining ranked lists of LCSHs, the weights for the same LCSH are added. As another example, the following table presents the top-ranked ten LCSHs (only subfield *a* in lower case) for the query *Vietnam War*.

| Rank | LCSH | Weight |
|---|---|---|
| 1 | world war, 1939-1945 | 16430.62 |
| 2 | vietnamese conflict, 1961-1975 | 15388.68 |
| 3 | united states | 13989.66 |
| 4 | world war, 1914-1918 | 8055.60 |
| 5 | vietnam | 6523.90 |
| 6 | war | 5503.86 |
| 7 | soldier | 2485.60 |
| 8 | great britain | 2209.91 |
| 9 | cold war | 2185.69 |
| 10 | world politics | 1725.20 |

## 5.3 LCSH to Words Entry Vocabulary Index

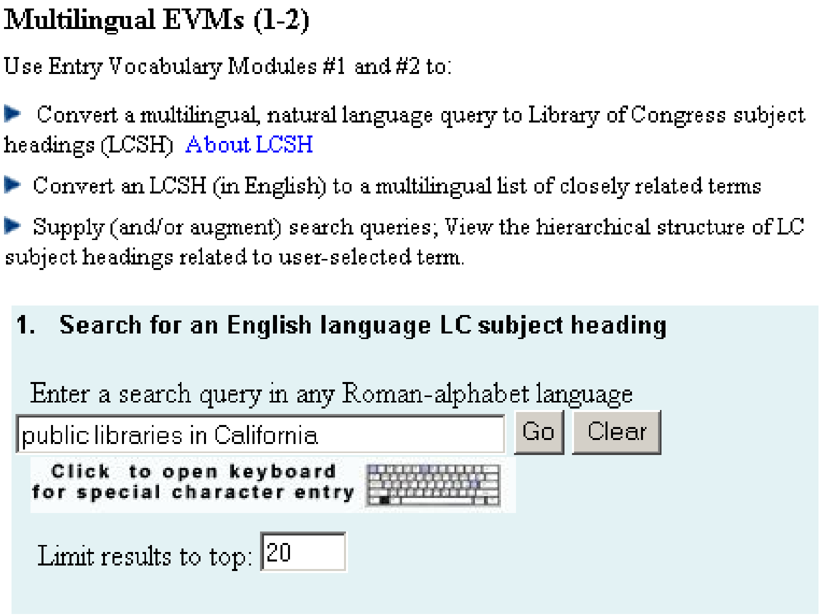
The examples presented in the previous section have demonstrated the use of association measures in creating entry vocabulary indexes that will prompt a ranked list of LCSHs that are most closely associated with a query. The same statistical association measure can also be used to create entry vocabulary indexes that will prompt a ranked list of *words* or *phrases* that are most closely associated with a LCSH. From the same training set of bibliographic records, we created a *LCSH-to-word* entry vocabulary index that returns a list of words that are most closely associated with a LCSH. As an example, the top-ranked twenty words, found in the title or notes fields, that are most closely associated with the subject heading *Alcoholism* are presented in the table below.

| Rank | LCSH (subfield *a*) | Weight |
|---|---|---|
| 1 | alcohol | 13471.94 |
| 2 | alcoholism | 11715.56 |
| 3 | abuse | 3708.09 |
| 4 | drug | 3467.22 |
| 5 | drink | 2563.53 |
| 6 | alcoholic | 2534.91 |
| 7 | treatment | 2349.03 |
| 8 | prevention | 1263.94 |
| 9 | problem | 1148.03 |
| 10 | addiction | 886.81 |
| 11 | alcoholismo | 865.75 |
| 12 | substance | 800.53 |
| 13 | alcoolisme | 436.38 |
| 14 | alkohol | 427.22 |
| 15 | dependence | 402.81 |
| 16 | drinker | 320.47 |
| 17 | alcool | 314.31 |
| 18 | addictive | 258.84 |
| 19 | recovery | 238.44 |
| 20 | programme | 231.75 |

Note the inclusion of foreign words (alcoholismo, alcoolisme, alkohol, and alcool). These words, derived from titles in foreign languages, demonstrate that the technique is language-independent and could be adopted in any country. It can also support diversity in U.S. libraries by allowing searches in Spanish or any other languages, so long as the training set contains content words in those languages in titles or abstracts. Both entry vocabulary indexes are publicly accessible at http://metadata.sims.berkeley.edu/prototypesI.html.

## 6 Access to an Online Catalog

To demonstrate the searching capability from a bibliographic record to some numeric database, the first step is to retrieve and display a bibliographic record from an online catalog. We implemented a web-based interface for searching online catalogs using an in-house implementation of the Z39.50 protocol. Besides the Z39.50 protocol, an important component that makes searching remote online catalogs feasible is the gateway between the HTTP (Hypertext Transfer Protocol) protocol and the Z39.50 protocol. While HTTP is a connectionless-oriented protocol, the Z39.50 is a connection-oriented protocol. The gateway maintains connections to remote Z39.50 servers. All search requests to any remote Z39.50 server go through the gateway.



Multilingual EVMs (1-2)

Use Entry Vocabulary Modules #1 and #2 to:

- Convert a multilingual, natural language query to Library of Congress subject headings (LCSH) About LCSH
- Convert an LCSH (in English) to a multilingual list of closely related terms
- Supply (and/or augment) search queries; View the hierarchical structure of LC subject headings related to user-selected term.

1. Search for an English language LC subject heading

Enter a search query in any Roman-alphabet language

public libraries in California

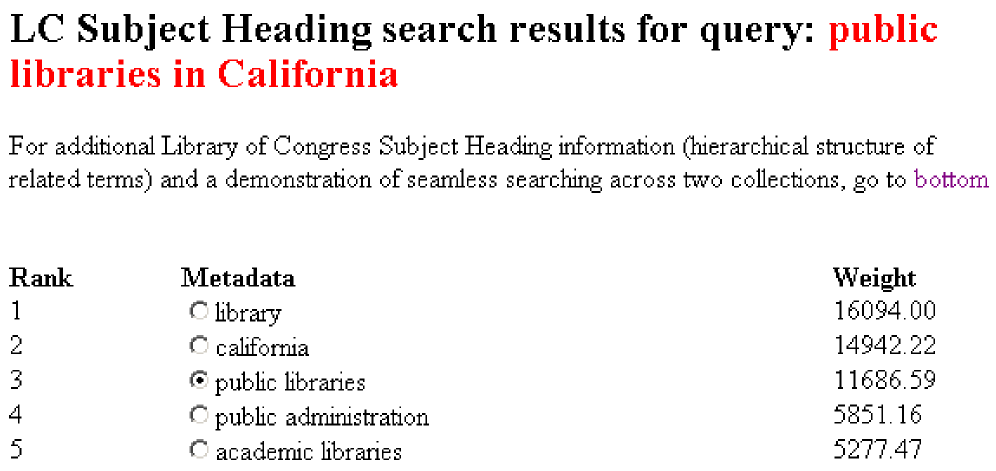
Click to open keyboard for special character entry

Limit results to top: 20

Figure 1: Search interface.

The web-based search interface, shown in Figure 1, takes a query as input and then submits the query to the *word-to-LCSH* entry vocabulary index. In response to the query, a ranked list of LCSHs that are most closely associated with the query is returned.

Figure 2 presents the top-ranked five subject headings most closely related to the query *public libraries in California*. From this ranked list of LCSHs, a user can select one LCSH, and then click on the *Search Melvyl* button to forward the selected LCSH as a query to the MELVYL online catalog. The web client will



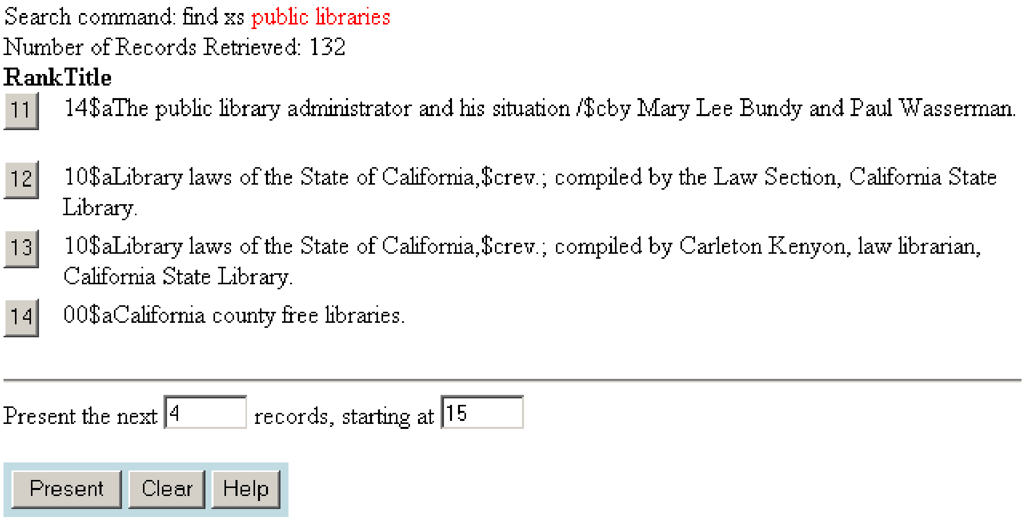
LC Subject Heading search results for query: public libraries in California

For additional Library of Congress Subject Heading information (hierarchical structure of related terms) and a demonstration of seamless searching across two collections, go to bottom

| Rank | Metadata | Weight |
|---|---|---|
| 1 | ○ library | 16094.00 |
| 2 | ○ california | 14942.22 |
| 3 | ● public libraries | 11686.59 |
| 4 | ○ public administration | 5851.16 |
| 5 | ○ academic libraries | 5277.47 |

Figure 2: EVI search results.

formulate a search query from the selected LCSH and send it to the web server which will, in turn, send it to the gateway. The gateway will establish a connection to the remote Z39.50 server (the MEYLVL online catalog) if the connection does not exist yet. The gateway will then send the search query to the remote Z39.50 server. An exact subject heading search will be executed at the MELVYL site. And the search results will be returned through the gateway and the web server to the web client for display on the user's screen. A search result entry, if any, contains only author's name and the title. The initial display may not list all the retrieved records since only a small number of bibliographic records is returned in response to each request due to limited message size. However, the user has the option to request more retrieved records by specifying the range of records to be returned using the Z39.50 protocol. Figure 3 presents 4 out of the



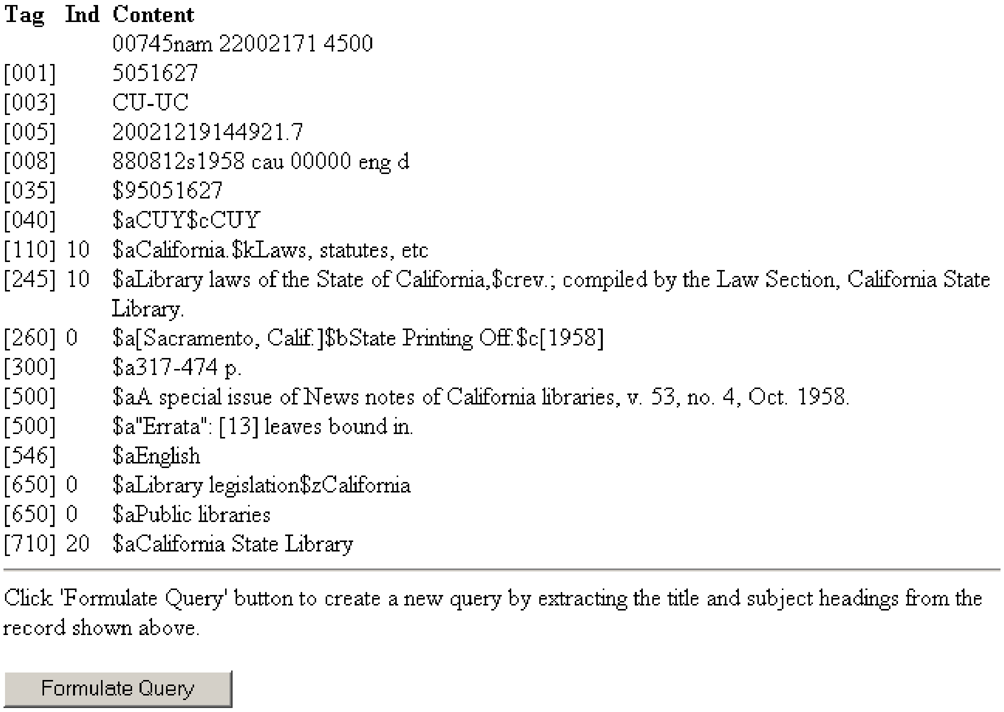
Search command: find xs public libraries
Number of Records Retrieved: 132

| Rank | Title |
|---|---|
| 11 | 14$aThe public library administrator and his situation /$cby Mary Lee Bundy and Paul Wasserman. |
| 12 | 10$aLibrary laws of the State of California,$crev.; compiled by the Law Section, California State Library. |
| 13 | 10$aLibrary laws of the State of California,$crev.; compiled by Carleton Kenyon, law librarian, California State Library. |
| 14 | 00$aCalifornia county free libraries. |

Present the next 4 records, starting at 15

Figure 3: Part of the search result returned by MELVYL.

132 retrieved titles as a result of performing an exact subject heading search on the query *public libraries* at the MELVYL site. From this list of search results displayed in short form, a user can choose to display the full MARC record in tagged form. The full MARC record in tagged form for title 12 is displayed in Figure 4.



| Tag | Ind | Content |
|---|---|---|
| | | 00745nam 22002171 4500 |
| [001] | | 5051627 |
| [003] | | CU-UC |
| [005] | | 20021219144921.7 |
| [008] | | 880812s1958 cau 00000 eng d |
| [035] | | $95051627 |
| [040] | | $aCUY$cCUY |
| [110] | 10 | $aCalifornia.$kLaws, statutes, etc |
| [245] | 10 | $aLibrary laws of the State of California,$crev.; compiled by the Law Section, California State Library. |
| [260] | 0 | $a[Sacramento, Calif.]$bState Printing Off.$c[1958] |
| [300] | | $a317-474 p. |
| [500] | | $aA special issue of News notes of California libraries, v. 53, no. 4, Oct. 1958. |
| [500] | | $a"Errata": [13] leaves bound in. |
| [546] | | $aEnglish |
| [650] | 0 | $aLibrary legislation$zCalifornia |
| [650] | 0 | $aPublic libraries |
| [710] | 20 | $aCalifornia State Library |

Click 'Formulate Query' button to create a new query by extracting the title and subject headings from the record shown above.

Figure 4: A MARC record displayed in tagged form.

# 7 Access to Numeric Databases

Creating an Entry Vocabulary Index requires a training set of records containing both descriptive words and topical metadata. This is often not readily available for numeric data sets. Our first effort was to create an Entry Vocabulary Index to the Standard Industrial Classification, widely used over many years in numeric data sets. This was feasible because we found a bibliographic database which used SIC codes to serve as a training set. But by the time the SIC Entry Vocabulary Index was completed, the SIC was discontinued, replaced by the North American Industry Classification System, so we created a mapping from SIC codes to NAICS codes. By this time it had become apparent that, with the current low level of interoperability in software and data formats, the labor required to create Entry Vocabulary Indexes and interfaces to numeric data sets was large. We could not deal with more than two or three data sets within the funding available, much less that we had hoped for. In order to prevent the costs of EVI and interface development from limiting the project in this way, we turned our attention to a collection of 3,000 numeric data sets available through a single interface and known as Counting California.

## 7.1 Numeric Database

The numeric database we used to demonstrate improved searching capabilities for accessing to numeric databases is the Counting California Numeric Database made available by California Digital Library at http://countingcalifornia.cdlib.org/. The database is a collection of some 3,000 numeric tables containing statistics related to a range of topics. The numeric datasets are mainly from the California Department of Health Services, California Department of Finance, and the Bureau of the Census. The tables are organized under a two-level classification scheme based on the topics. There are 16 topics at the top level, which are subdivided into 184 subtopics (Source: http://countingcalifornia.cdlib.org/). All the numeric tables are placed under the subtopics, some may placed under more than one subtopic. The top level topics are:

1 Agriculture and Natural Resources
2 Banking, Finance and Insurance
3 Business and Industry
4 Crime, Law Enforcement & Criminal Justice
5 Education
6 Elections
7 Energy and Public Utilities
8 Health and Vital Statistics
9 Housing Characteristics and Costs
10 Income, Poverty, and Cost of Living
11 Land, Water and Climate
12 Population and Demographics
13 Public Finance, Government & Taxes
14 Social Services and Public Assistance
15 Transportation
16 Work, Labor, and Employment

Source: http://countingcalifornia.cdlib.org/ (2002)

As an example, the subtopics under the main topic *Agriculture and Natural Resources* are:

1 Farms and Farming
2 Fishing
3 Forestry and Lumber
4 Minerals

At the Counting California website, a user can browse the tables by topic, starting from a top level topic, to a selected subtopic, then to a selected table. The Boolean searching of the tables is also provided.

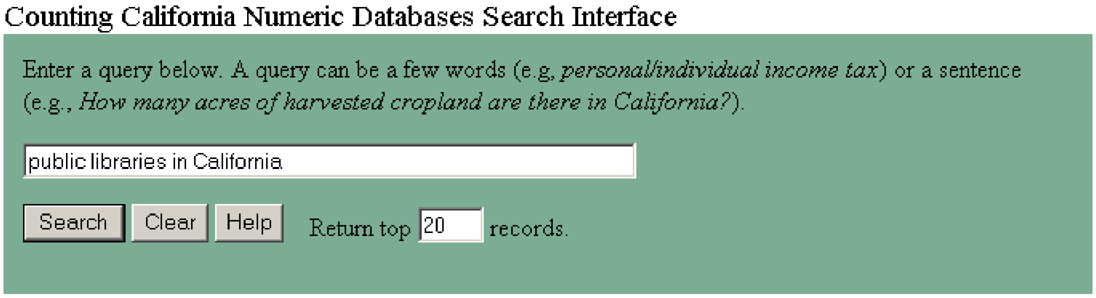
We have provided two new ways to access to this set of numeric tables, *probabilistic retrieval* and *EVI-based retrieval*. We have extracted the captions of some 3,000 tables from the Counting California website at http://countingcalifornia.cdlib.org/, and treated the caption of a table as a record. An extracted sample record is shown in the following table.

```
<table>
<topic> education </topic>
<subtopic> libraries </subtopic>
<caption>
LIBRARY STATISTICS, STATEWIDE SUMMARY BY TYPE OF LIBRARY CALIFORNIA,
1992-93 TO 1997-98
</caption>
</table>
```

Each record contains the caption of a numeric table, the subtopic under which the table is placed, and topic at the top level.

## 7.2 Probabilistic Access to a Numeric Database

We created a word index for this collection of about 3,000 records. The texts in the *caption* field were tokenized first; stopwords removed; and then the content words normalized. We provided an web-based search interface, shown in Figure 5 and available at http://otlet.sims.berkeley.edu/countingcalifornia.html, that will take queries in free form. A query can be a single word, a phrase, a set of keywords, incomplete or complete sentences.



Counting California Numeric Databases Search Interface

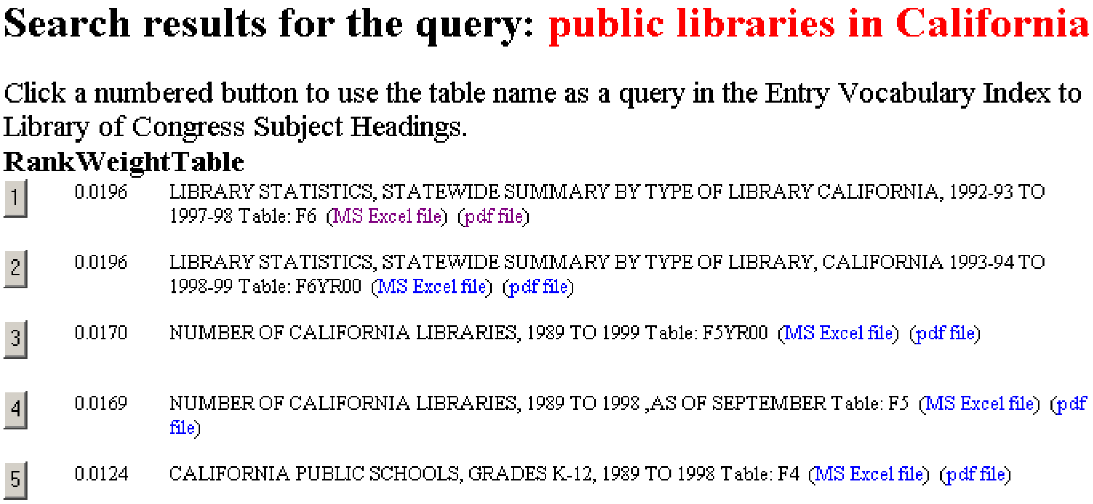
Enter a query below. A query can be a few words (e.g, *personal/individual income tax*) or a sentence (e.g., *How many acres of harvested cropland are there in California?*).

public libraries in California

Return top 20 records.

Figure 5: Search interface for Counting California numeric databases.

Our search engine uses an in-house implementation of a probabilistic full-text retrieval algorithm developed at Berkeley. Details on the retrieval algorithm can be found in [1]. The search engine takes a free form query and returns a ranked list of captions of the tables ranked according to their relevance scores. The more likely relevant tables with respect to the query are ranked higher than those that are less likely. As an example, the top-ranked 5 captions returned by our search engine in response to the query “public libraries in California” are shown in Figure 6.



**Search results for the query: public libraries in California**

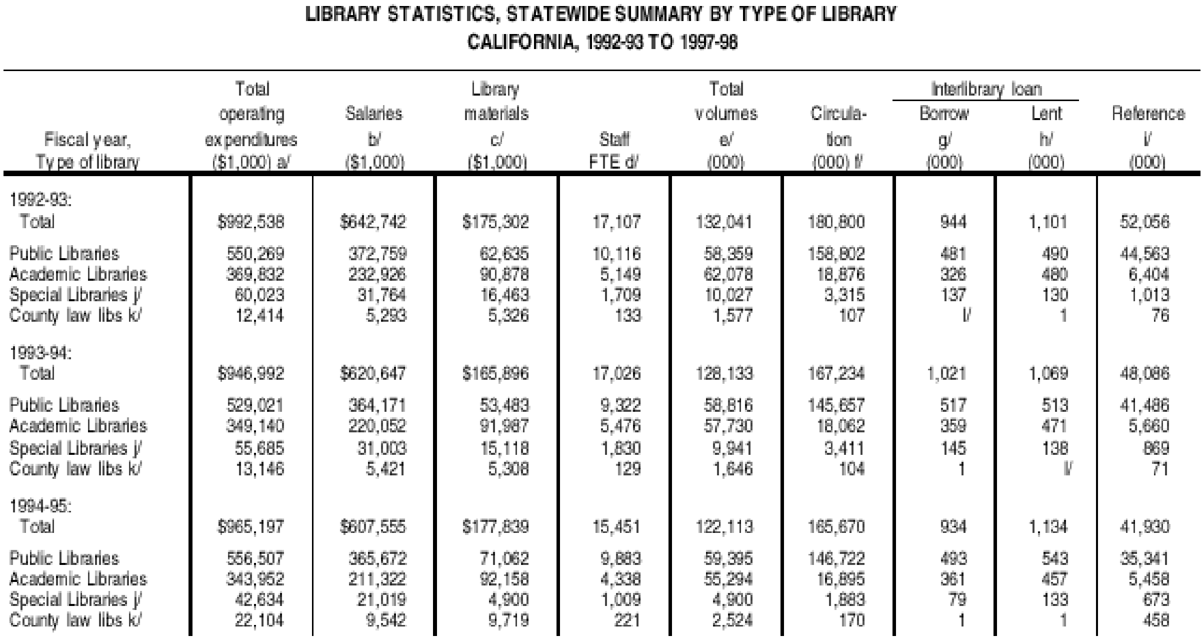
Click a numbered button to use the table name as a query in the Entry Vocabulary Index to Library of Congress Subject Headings.

| Rank | Weight | Table |
|---|---|---|
| 1 | 0.0196 | LIBRARY STATISTICS, STATEWIDE SUMMARY BY TYPE OF LIBRARY CALIFORNIA, 1992-93 TO 1997-98 Table: F6 (MS Excel file) (pdf file) |
| 2 | 0.0196 | LIBRARY STATISTICS, STATEWIDE SUMMARY BY TYPE OF LIBRARY, CALIFORNIA 1993-94 TO 1998-99 Table: F6YR00 (MS Excel file) (pdf file) |
| 3 | 0.0170 | NUMBER OF CALIFORNIA LIBRARIES, 1989 TO 1999 Table: F5YR00 (MS Excel file) (pdf file) |
| 4 | 0.0169 | NUMBER OF CALIFORNIA LIBRARIES, 1989 TO 1998 ,AS OF SEPTEMBER Table: F5 (MS Excel file) (pdf file) |
| 5 | 0.0124 | CALIFORNIA PUBLIC SCHOOLS, GRADES K-12, 1989 TO 1998 Table: F4 (MS Excel file) (pdf file) |

Figure 6: Search results in the Counting California database for the query *public libraries in California*.

This search engine has a couple of advantages over commonly used Boolean retrieval. Firstly, it takes queries in free form. Secondly, the results are ranked according to their relevance to the query. The retrieval algorithm has been tested on ten languages, including Arabic, Chinese, and Spanish, in three large-scale text retrieval evaluations conferences: TREC, CLEF, and NTCIR. The retrieval algorithm has been shown effective for all the languages that have been tested.

Each entry in the result list is linked to a numeric table maintained at the Counting California website. By clicking on the appropriate link, a user can display the numeric table in PDF format or in MS Excel format. Figure 7 displays part of the numeric table top-ranked in Figure 6.



LIBRARY STATISTICS, STATEWIDE SUMMARY BY TYPE OF LIBRARY
CALIFORNIA, 1992-93 TO 1997-98

| | | | | | | | Interlibrary loan | | |
|---|---|---|---|---|---|---|---|---|---|
| Fiscal year, Type of library | Total operating expenditures ($1,000) a/ | Salaries b/ ($1,000) | Library materials c/ ($1,000) | Staff FTE d/ | Total volumes e/ (000) | Circulation (000) f/ | Borrow g/ (000) | Lent h/ (000) | Reference i/ (000) |
| 1992-93: | | | | | | | | | |
| Total | $992,538 | $642,742 | $175,302 | 17,107 | 132,041 | 180,800 | 944 | 1,101 | 52,056 |
| Public Libraries | 550,269 | 372,759 | 62,635 | 10,116 | 58,359 | 158,802 | 481 | 490 | 44,563 |
| Academic Libraries | 369,832 | 232,926 | 90,878 | 5,149 | 62,078 | 18,876 | 326 | 480 | 6,404 |
| Special Libraries j/ | 60,023 | 31,764 | 16,463 | 1,709 | 10,027 | 3,315 | 137 | 130 | 1,013 |
| County law libs k/ | 12,414 | 5,293 | 5,326 | 133 | 1,577 | 107 | l/ | 1 | 76 |
| 1993-94: | | | | | | | | | |
| Total | $946,992 | $620,647 | $165,896 | 17,026 | 128,133 | 167,234 | 1,021 | 1,069 | 48,086 |
| Public Libraries | 529,021 | 364,171 | 53,483 | 9,322 | 58,816 | 145,657 | 517 | 513 | 41,486 |
| Academic Libraries | 349,140 | 220,052 | 91,987 | 5,476 | 57,730 | 18,062 | 359 | 471 | 5,660 |
| Special Libraries j/ | 55,685 | 31,003 | 15,118 | 1,830 | 9,941 | 3,411 | 145 | 138 | 869 |
| County law libs k/ | 13,146 | 5,421 | 5,308 | 129 | 1,646 | 104 | 1 | l/ | 71 |
| 1994-95: | | | | | | | | | |
| Total | $965,197 | $607,555 | $177,839 | 15,451 | 122,113 | 165,670 | 934 | 1,134 | 41,930 |
| Public Libraries | 556,507 | 365,672 | 71,062 | 9,883 | 59,395 | 146,722 | 493 | 543 | 35,341 |
| Academic Libraries | 343,952 | 211,322 | 92,158 | 4,338 | 55,294 | 16,895 | 361 | 457 | 5,458 |
| Special Libraries j/ | 42,634 | 21,019 | 4,900 | 1,009 | 4,900 | 1,883 | 79 | 133 | 673 |
| County law libs k/ | 22,104 | 9,542 | 9,719 | 221 | 2,524 | 170 | 1 | 1 | 458 |

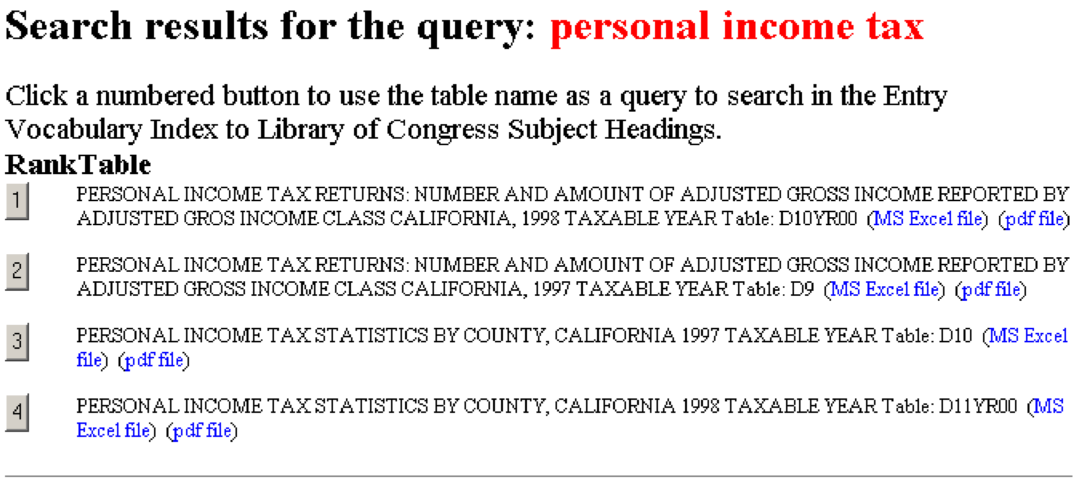
Figure 7: Part of a numeric table (source: http://countingcalifornia.cdlib.org/).

### 7.3 Entry Vocabulary Index-based Access to a Numeric Database

From the extracted records, we created a word to subtopics entry vocabulary index. The words are extracted from the captions of the tables, and the subtopics from the topic classification scheme developed by the Counting California Project. A web interface is provided at http://otlet.sims.berkeley.edu/countingcaliforniaEVI.html. As an example, the subtopics ranked in the top ten for the query "personal/individual income tax" are:

| rank | subtopic | weight |
|---|---|---|
| 1 | income | 542.53 |
| 2 | government earnings and tax revenues | 251.71 |
| 3 | personal income | 156.67 |
| 4 | property tax | 74.58 |
| 5 | personal income tax | 59.99 |
| 6 | corporate income tax | 56.84 |
| 7 | per capita income | 44.25 |
| 8 | population | 43.46 |
| 9 | sales tax | 35.92 |
| 10 | age | 26.94 |

A user can click on a selected subtopic in the ranked lists of subtopics to view the captions of all the tables that are classified under the chosen subtopic. Clicking on "personal income tax" produced the list of captions shown in Figure 8.



Search results for the query: personal income tax

Click a numbered button to use the table name as a query to search in the Entry Vocabulary Index to Library of Congress Subject Headings.

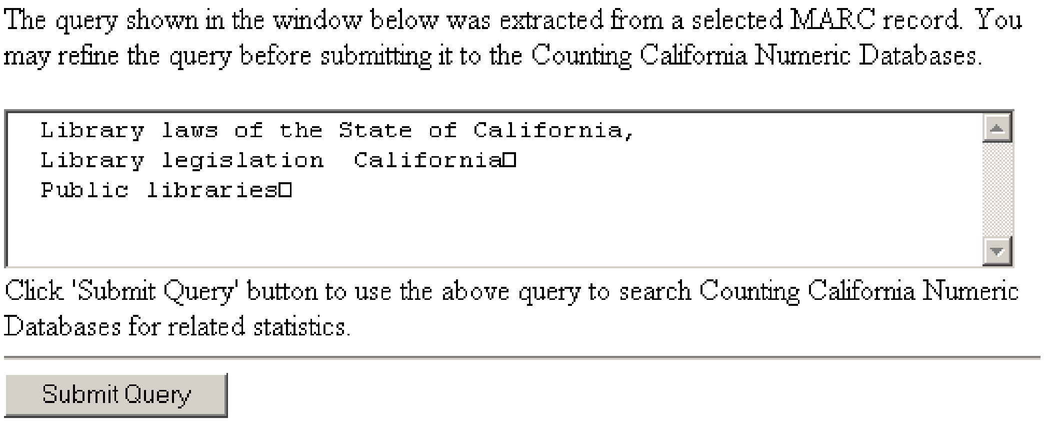
| Rank | Table |
|---|---|
| 1 | PERSONAL INCOME TAX RETURNS: NUMBER AND AMOUNT OF ADJUSTED GROSS INCOME REPORTED BY ADJUSTED GROS INCOME CLASS CALIFORNIA, 1998 TAXABLE YEAR Table: D10YR00 (MS Excel file) (pdf file) |
| 2 | PERSONAL INCOME TAX RETURNS: NUMBER AND AMOUNT OF ADJUSTED GROSS INCOME REPORTED BY ADJUSTED GROSS INCOME CLASS CALIFORNIA, 1997 TAXABLE YEAR Table: D9 (MS Excel file) (pdf file) |
| 3 | PERSONAL INCOME TAX STATISTICS BY COUNTY, CALIFORNIA 1997 TAXABLE YEAR Table: D10 (MS Excel file) (pdf file) |
| 4 | PERSONAL INCOME TAX STATISTICS BY COUNTY, CALIFORNIA 1998 TAXABLE YEAR Table: D11YR00 (MS Excel file) (pdf file) |

Figure 8: The list of tables placed under the subtopic *personal income tax*.

# 8 Traverse Searching Between Online Catalogs and Number Databases

## 8.1 Searching Numeric Databases from Bibliographic Records

In section 6 we talked about how to access to an online catalog through a word-to-LCSH entry vocabulary index, and display a full MARC record in tagged form. If the user wishes to know if there are any interesting facts or statistics in a numeric database that are related to the same topic revealed in the displayed marc record, he/she can click the *formulate query* button placed at the end of a displayed full MARC record to create a query for searching a numeric database. The initial query will contain the texts extracted from the subfields *a* and *b* of the title field 245 and the subject headings in the displayed full MARC record. The initial query is placed in a new window where the user can subsequently refine the query before submitting it to the search engine for a numeric database. Figure 9 shows the query extracted from the MARC record displayed in Figure 4. The search engine returns a ranked list of captions of the tables that are most likely relevant to the query. From the ranked list of captions displayed, a user can choose to view the full table either in PDF format or MS Excel format. Figure 10 shows the search results in the Counting California database using the extracted query displayed in Figure 9.

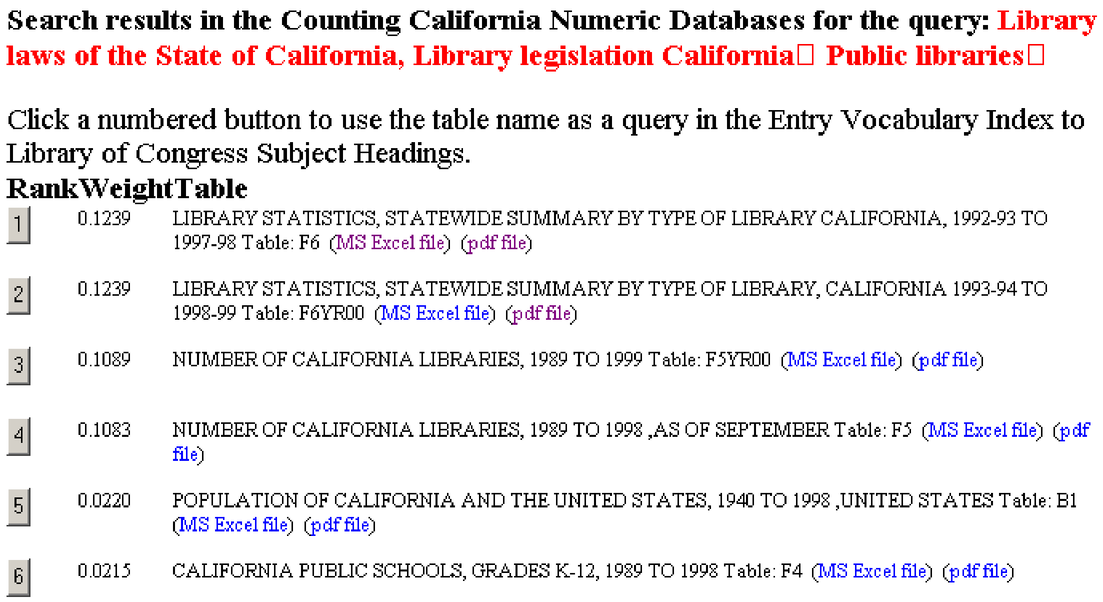


The query shown in the window below was extracted from a selected MARC record. You may refine the query before submitting it to the Counting California Numeric Databases.

```
Library laws of the State of California,
Library legislation  California□
Public libraries□
```

Click 'Submit Query' button to use the above query to search Counting California Numeric Databases for related statistics.

Figure 9: Query extracted from a MARC record.



**Search results in the Counting California Numeric Databases for the query: Library laws of the State of California, Library legislation California□ Public libraries□**

Click a numbered button to use the table name as a query in the Entry Vocabulary Index to Library of Congress Subject Headings.

| Rank | Weight | Table |
|---|---|---|
| 1 | 0.1239 | LIBRARY STATISTICS, STATEWIDE SUMMARY BY TYPE OF LIBRARY CALIFORNIA, 1992-93 TO 1997-98 Table: F6 (MS Excel file) (pdf file) |
| 2 | 0.1239 | LIBRARY STATISTICS, STATEWIDE SUMMARY BY TYPE OF LIBRARY, CALIFORNIA 1993-94 TO 1998-99 Table: F6YR00 (MS Excel file) (pdf file) |
| 3 | 0.1089 | NUMBER OF CALIFORNIA LIBRARIES, 1989 TO 1999 Table: F5YR00 (MS Excel file) (pdf file) |
| 4 | 0.1083 | NUMBER OF CALIFORNIA LIBRARIES, 1989 TO 1998 ,AS OF SEPTEMBER Table: F5 (MS Excel file) (pdf file) |
| 5 | 0.0220 | POPULATION OF CALIFORNIA AND THE UNITED STATES, 1940 TO 1998 ,UNITED STATES Table: B1 (MS Excel file) (pdf file) |
| 6 | 0.0215 | CALIFORNIA PUBLIC SCHOOLS, GRADES K-12, 1989 TO 1998 Table: F4 (MS Excel file) (pdf file) |

Figure 10: Search results in the Counting California database using the query extracted from a MARC record.

## 8.2 Searching Online Catalogs from Numeric Tables

If a user started with searching the Counting California numeric database for a topic, and is interested in literature on the same topic in the online catalog, the user can click the numbered button placed at the beginning of each result entry. After clicking a numbered button in the search results, the caption associated with the numbered button will be forwarded as a query to the word-to-LCSH entry vocabulary index. Clicking the numbered button "1" in Figure 6 resulted in submitting the associated caption as a query to the *word-to-LCSH* entry vocabulary index. The top-ranked seven subject headings that are most closely associated with the selected caption used as query are shown in Figure 11. The process of viewing a full MARC record is the same as that described in section 6.

## 8.3 Implementation

Figure 12 presents the diagram showing an implementation of seamless search of numeric and bibliographic/textual resources. The boxes shown in the figure are:

1. A search interface, shown in Figure 1, for accessing bibliographic/textual resources through a word-to-LCSH entry vocabulary index.

2. A word to the LCSH entry vocabulary index.

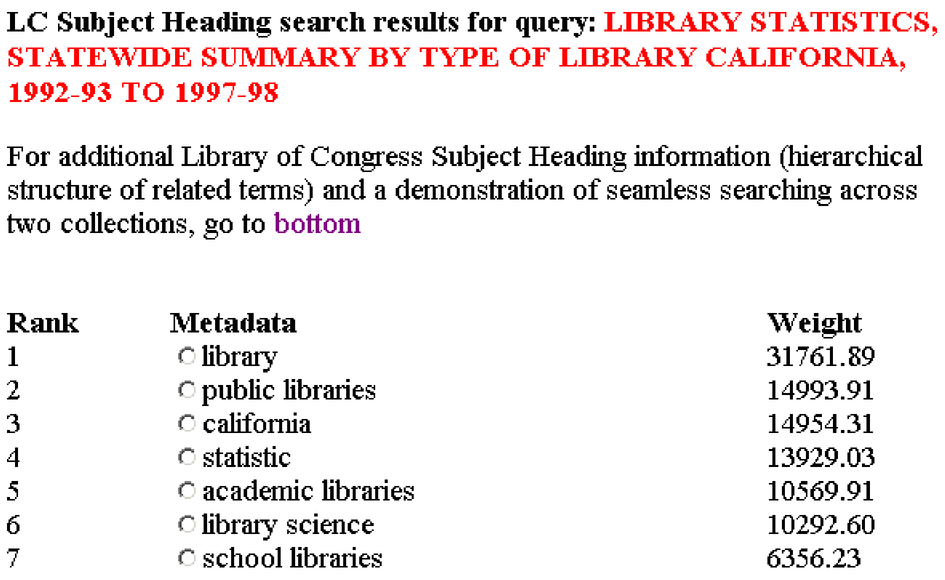
3. A ranked list of LCSHs closely associated with the query, as in Figure 2 or 11.

4. An online catalog.

5. Results of searching the online catalog using a LCSH, as in Figure 3.

6. A full MARC record displayed in tagged form, as in Figure 4.

7. A new query formed by extracting the title and subject fields from the displayed full marc record, as in Figure 9.



**LC Subject Heading search results for query: LIBRARY STATISTICS, STATEWIDE SUMMARY BY TYPE OF LIBRARY CALIFORNIA, 1992-93 TO 1997-98**

For additional Library of Congress Subject Heading information (hierarchical structure of related terms) and a demonstration of seamless searching across two collections, go to bottom

| Rank | Metadata | Weight |
|---|---|---|
| 1 | ○ library | 31761.89 |
| 2 | ○ public libraries | 14993.91 |
| 3 | ○ california | 14954.31 |
| 4 | ○ statistic | 13929.03 |
| 5 | ○ academic libraries | 10569.91 |
| 6 | ○ library science | 10292.60 |
| 7 | ○ school libraries | 6356.23 |

Figure 11: The search results in the LCSH entry vocabulary index for the query extracted from a table caption.

8. A numeric database.

9. A list of captions of numeric tables ranked by relevance score to the query, as in Figure 6.

10. Numeric table displayed in PDF or MS Excel format, as in Figure 7.

11. A search interface, shown in Figure 5, for numeric databases based on a probabilistic search algorithm.

A user can start a search using either interface, and find records on the same topic of interest in bibliographic/textual databases and socio-economic databases.

# 9 Future Work

## 9.1 Geographical access to numeric databases

Socio-economic numeric data sets nearly always have a geographical aspect: the data refer to particular places or areas and searchers very commonly want data pertaining to a place. We found that this was hard to achieve for several reasons. Place names are ambiguous and unstable: A search for data relating to Trinidad might lead to Trinidad, West Indies, instead of Trinidad, California, for example. With numeric databases the problem is compounded because specialized geo-political divisions, such as census tracts and counties, are used. These divisions do not match conveniently with searchers' use of place-names. Eventually we concluded that reliance on the names of places could never work satisfactorily. The only effective path to reliable access to data relating to places would be to use geo-spatial coordinates (latitude and longitude) to establish unambiguously the identity and location of any place and the relationship between places. Data relating to Berkeley may be available only in aggregated data for Alameda County. This means that gazetteers and map visualizations become important. Gazetteers relate place names to locations, locations to place names, and reveal spatial relationships between places, e.g. the city of Alameda is an island within Alameda County. It was this problem that led us to propose the recently approved IMLS National Library Leadership Award entitled "Going Places in the Catalog: Improved Geographical Access."

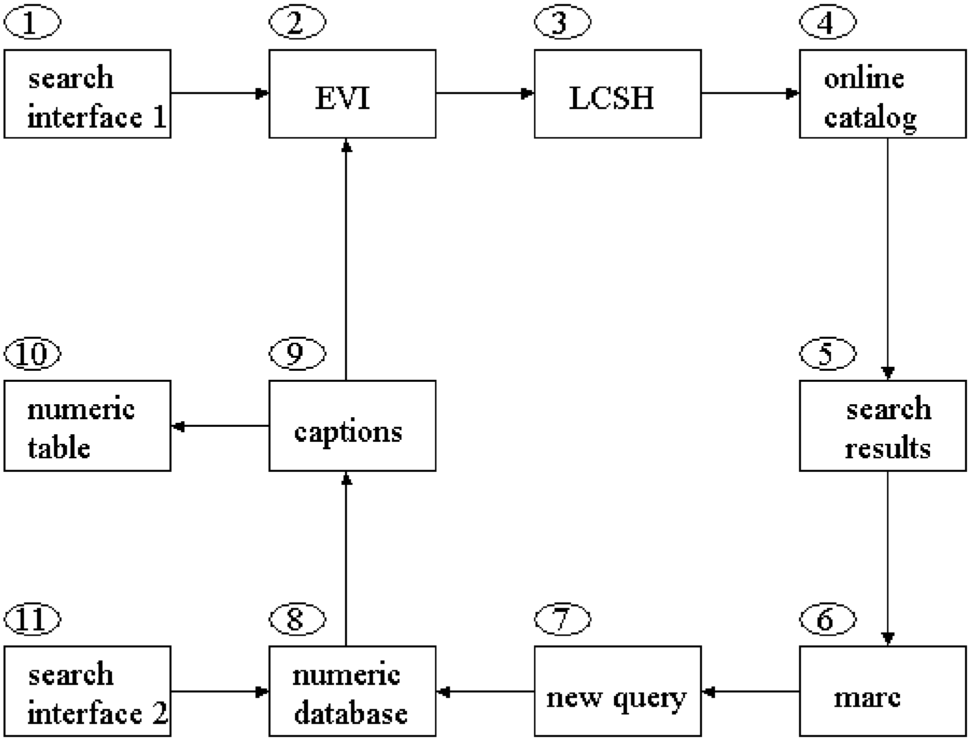


Figure 12: Implementation of seamless searching of numeric (socio-economic) and bibliographic/textual resources.

## 9.2 Enhanced access to numeric databases

The descriptive texts, such as the caption, associated with numeric tables are often brief and concise, which provides a rather limited basis for locating the table in response to queries. Sometimes the caption/title of a table may be the only textual description about the content of the table that is searchable. The titles are sometimes very general. For example, one of the tables in Counting California databases is named “LIBRARY STATISTICS, STATEWIDE SUMMARY BY TYPE OF LIBRARY CALIFORNIA, 1992-93 TO 1997-98.” The title is so general that neither the kinds of statistics nor the types of libraries are revealed in the title. If a user poses the question “What is the total operating expenditures of public libraries in California?” to a query system that indexes table titles only, the search may well be ineffective since the only word in common between the table title and the user’s query is ‘library’, assuming the plurals of the nouns are normalized to singular form.

The table column headings and row headings provide additional information about the content of a numeric table. However, the column and row headings are usually not directly searchable. For example, a table named “Language spoken at home” in Counting California databases consists of rows and columns. The column headings list the languages spoken at home, while the row headings show the county names in California. Each cell in the table gives the number of people, 5 years old or over, who speak a specific language at home. To answer questions, such as “How many people speak Spanish at home in Alameda County, California?”, using the table title alone may not retrieve the table that contains the answer to the example question. If the textual description were enriched by including the column and row headings, the chances of retrieving the tables of interest should be improved.

We recommend that the textual descriptions of a numeric table be enriched. This could be done automatically by combining the table title and its column and row headings.

## 10 Administration

Research and demonstration projects do not always develop as intended. This was the case with the present project. After the proposal for this project had been submitted, the P.I.s received substantial funding from the Defense Advanced Research Projects Agency for research on the creation and evaluation of Entry Vocabulary Indexes. This additional funding benefited our work for IMLS because it enabled us to understand better what we needed to do and also to buy faster computers and the large amounts of disk storage without which we could not have handled millions of MARC records. It also provided travel funds that enabled us to attend more conferences and to explain our IMLS-related work to a wider audience. In the circumstances, as explained in our informal six-monthly reports, it made sense to slow the pace of the IMLS project in order to reap for it some of the benefit from the DARPA-funded research. This was why we requested a no-cost extension and stretched a two year project over three years.

An Advisory Committee was appointed and we are very grateful to George McGregor (Chiron Corporation, Emeryville, CA), Gary Peete (Haas Business School Library, University of California, Berkeley), Vivian Pisano (San Francisco Public Library), and Andrea Sevetson (Government Documents Specialist, University Library, UC Berkeley) for agreeing to serve. The purpose for which we most wanted the Advisory Committee was to advise on usability of the interface design and on the selection of databases to link to. In the event, the technical difficulties of making the promised prototype work at all meant that we did not reach the stage at which we could explore these refinements and, as a result, little use was made of the Advisory Committee.

The project website is at http://metadata.sims.berkeley.edu/GrantSupported/seamless.html. Publications describing the work completed are in preparation and a list of related publications is given below. Financial reports are being submitted separately by the campus central accounting staff.

Dr Aitao Chen played a central role in the conduct of this project. Graduate Student Assistants Hui-Min Chen, Michael Gebbie, Karthik Gourisankaran, Natalia Perelman, Joanna Plattner, and Jacek Purat also worked on it.

The work that we undertook in this project proved more difficult than expected. In particular, we found that building access to numeric data sets was harder and more time-consuming than expected. As standards evolve, interoperability will be easier. We were, we like to think, ahead of our time. Nevertheless we completed the tasks we undertook to do.

## 11 Related Publications

Michael Buckland.
Entry Vocabulary, Intermediaries, and Retrieval Performance. In: *Information in a Networked World: Harnessing the Flow. Proceedings of the 64th Meeting of the American Society for Information Science and Technology*, Nov 3-8, 2001, Washington, DC. Medford, NJ: Information Today, 2001. pp 112-117.

Fredric C. Gey, Michael Buckland, Aitao Chen, and Ray Larson.
Entry Vocabulary – a Technology to Enhance Digital Search. In: *Proceedings of the First International Conference on Human Language Technology*, San Diego, March 2001, pp 91-95.
http://metadata.sims.berkeley.edu/papers/hlt01-final.pdf

Youngin Kim, Barbara Norgard, Aitao Chen, Fredric Gey.
Using Ordinary Language to Access Metadata of Diverse Types of Information Resources: Trade Classification and Numeric Data. In: *Knowledge: Creation, Organization, and Use. Proceedings of the American*

*Society for Information Science Annual Meeting*, Oct 31-Nov 4, 1999, Washington DC. Medford, NJ: Information Today, 1999, pp. 172-180.

## References

[1] William S. Cooper, Aitao Chen, and Fredric C. Gey. Full text retrieval based on probabilistic equations with coefficients fitted by logistic regression. In D. K. Harman, editor, *The Second Text REtrieval Conference (TREC-2)*, pages 57–66, March 1994.

[2] Martin Zaidel Daniel Karp, Yves Schabes and Dania Egedi. A freely available wide coverage morphological analyzer for english. In *Proceedings of COLING*, 1992.

[3] Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19:61–74, March 1993.