

Submitted for publication *Information Technology and Libraries*. Forthcoming December 2006.
See <http://www.ala.org/ala/lita/litapublications/ital/252006/vol252006.htm> Published version
may differ a little.

Search Across Different Media: Numeric Data Sets and Text Files

Rev. July 18, 2006.

Michael Buckland,
Emeritus Professor, School of Information, University of California, Berkeley, CA 94720-4600

Aitao Chen,
Yahoo! Inc., Sunnyvale, CA 94089

Fredric C. Gey,
Information Scientist, UC Data Archive & Technical Assistance, University of California,
Berkeley, CA 94720-5100

Ray R. Larson
Professor, School of Information, University of California, Berkeley, CA 94720-4600

Correspondence to: buckland@sims.berkeley.edu

ACKNOWLEDGMENT

This work was partially supported by the Institute of Museum and Library Services through National Library Leadership Grant No 178 for a project entitled “Seamless Searching of Numeric and Textual Resources,” and was based on prior research partially supported by DARPA Contracts N66001-97-C-8541; AO# F477: “Search Support for Unfamiliar Metadata Vocabularies” and N66001-00-1-8911, TO# J290: “Translingual Information Management Using Domain Ontologies.”

ABSTRACT

Digital technology encourages hope of searching across and between different media forms (text, sound, image, numeric data). We describe topic searches in two different media: text files and socio-economic numeric databases and also for transverse searching, whereby retrieved text is used to find topically related numeric data and vice versa. Direct transverse searching across different media is impossible. Descriptive metadata provides enabling infrastructure, but usually requires mappings between different vocabularies and a search term recommender system. Statistical association techniques and natural language processing can help. Searches in socio-economic numeric databases ordinarily require that place and time be specified.

INTRODUCTION

A hope for libraries is that new technology will support searching across an increasing range of resources in a growing digital landscape. The rise of the Internet provides a technological basis for shared access to a very wide range of resources. 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. However, the use of digital technology implies a degree of technical compatibility between different media, sometimes referred to as “media convergence,” and these developments encourage the prospect of being able to search across and between different media forms, notably text, images, sound, and numeric data sets, for different kinds of material relating to the same topic. To examine the practical problems involved, we undertook to demonstrate searching between and across two different media forms: text files and socio-economic numeric data sets. (Additional technical details of this work and screen shots can be found in ¹).

Two kinds of search are needed: First, it should be possible to do a topical search in multiple media resources, so that one can find, for example, both pertinent factual numeric data and relevant discussion. One difficulty is that the vocabulary used to classify the numeric data is ordinarily quite different from the subject headings used for books, magazine articles, and newspaper stories about the same topic. Second, when an intriguing data value is encountered, one would like to move directly to topically relevant texts. Likewise, third, when a questionable statement is read, one would like to be able to find relevant statistical evidence. There needs, therefore, to be search support that facilitates such transverse searching between resources, establishing connections, transferring data, and invoking appropriate utilities in a helpful way.

We addressed both problems through the design and demonstration of a gateway providing search support for both text and socio-economic numeric databases. First, the gateway should help users conduct searches in databases of different media forms by accepting a query in the searcher’s own terms and then suggesting the specialized categorization terms to search for in the selected resource. Second, if you found something interesting in a socio-economic database, the gateway would help you to find documents on the same topic in a text database, and vice versa. Selection of the best search terms in target databases is supported by the use of indexes to the categories (entries, headings, class numbers, etc.) in the system to be searched. These search term recommender systems (also known as “entry vocabulary indexes”) resemble Melvil Dewey’s “Relativ Index,” but are created using statistical association techniques.²

Four characteristics of our investigation need to be noted:

1. Searching independent sources: We were not concerned with ingesting resources from different sources into a consolidated local data repository and searching within it. Instead, we were interested in being able to search effectively in any accessible resource as and when one wants. This implies that interoperability issues in dealing with the native query languages and metadata vocabularies of remote repositories can be solved.
2. Search for independent content: Numeric data sets commonly have associated text in the form of documentation, code books, and commentary. However, we were interested in finding topical content that was had no such formal or literary connection. By “independent” we mean, for

example, a newspaper article written by someone unaware that relevant statistical data existed or written before they existed. In the other direction, having found statistical data of interest, could we find topically related text created independently of this particular data point?

3. We chose two different media forms: text and numeric data sets. They look similar because they both use arabic numerals, but the traditional reliance in information retrieval in a text environment of using any character string from the corpus as a query, although technically feasible, cannot be expected to be useful here. One can copy a number expressing quantity, such as 12,941, from a numeric data cell, use it as a query in a text search engine such as Google and retrieve a large and eclectic retrieved set, usually involving 12941 as an identifying number for a postal code, a memorandum, a part number, software bug report, and so on, but the relationship is spurious. It requires great faith in numerology to expect anything topically meaningful to the original data cell one started with. With other combinations of media forms not even spurious results are feasible: one cannot submit a musical fragment or some pixels from an image as a text query.

4. Our interest was in how to achieve a better return on existing investments in well-formed, edited resources with descriptive metadata. This project built directly on prior work on how to make more effective use of existing, expertly developed metadata, rather than creating or replacing metadata.

Search of multiple resources comes in two forms:

1. *Parallel search* is when a single query is sent to two or more resources at more or less the same time. For example, a researcher interested in the import of shrimps would like to see both articles in newspapers and also trade statistics. So one might send a query to the Census Bureau's U.S. Imports and Exports numeric data series and look at SIC 0913 for shrimp and prawn and note a dramatic increase in imports from Vietnam through Los Angeles from 1995 onwards. One would also search newspaper indexes for articles such as "Normalizing ties to Vietnam important steps for U.S. firms; California stands to profit handsomely when barriers fall to trade with fast-growing country" (*Los Angeles Times* July 12, 1995:D1). Different sources are likely to use different index terms or categories, so the challenge is how to express the searcher's query in terms that will be effective for searching in the target resources, which, mostly likely, will use different vocabularies. As one example, the term for "automobiles" is 3711 in the Standard Industrial Classification; TL 205 in the Library of Congress Classification, 180/280 in the U.S. Patent Classification; and, in the Census Bureau's U.S. Imports and Exports data series, PASS MOT VEH, SPARK IGN ENG.³

2. *Transverse search* is when an item of interest found in one resource is used as the basis for a query to be forwarded to a different resource. The challenge here is, again, that when a query using the topical metadata in one resource needs to be expressed in the vocabulary of the target resource, the metadata vocabularies in the two resources will usually be different from each other, and, quite likely, both are unfamiliar for the searcher.

When searching within a single media form, it may be possible to use content itself

directly as a query: A fragment of text in a source text database is commonly used to as query in a target text database. Similarly one might start with an image and seek images that are measurably similar. But because such direct search cannot be done when searching across different media forms. An indirect approach relying on the use of interpretive representations becomes necessary. As the network environment expands, mapping *between* vocabularies increasingly important.

TEXT AND NUMERIC RESOURCES

Text resource

As a text file we chose to use a library catalog, a special case of text file, rather than a corpus of “full text.” The reasons were practical: In this exploratory investigation, we needed to start with resources that had rich metadata; we wanted a resource that was sufficiently under our control for us to be able to experiment with it; a library catalog was in the spirit of the project in that it would lead to additional text resources; and we had a suitable resource, which we intended to use anyway for metadata mapping: a set of several million MARC records, derived from MELVYL, the University of California online library catalog.

Socio-economic Numeric Data Set

Initially, and in prior work, we had worked on access to U.S. federal data series, especially import and export statistics and county business reports. Although we made some progress with interfaces to these data series, it became clear that the investment needed to craft interoperable access was high relative to the staff available to us. Crafting access to individual data series did not appear to be a scalable way to demonstrate variety within our limited resources, so we turned our attention to a single collection comprising many diverse numeric tables, the *Counting California* database.⁴

MAPPING TOPICAL METADATA

Well-edited, high quality databases typically have topical metadata expertly assigned from a vocabulary (thesaurus, classification, subject heading system, or set of categories). But there is a Babel of different vocabularies. Not only do the names of topics vary, but the underlying concepts or categories may also differ. Effective searching requires expert familiarity with a system’s vocabulary, but, as access to digital resources expands, the diversity of vocabularies increases and accessible resources are decreasingly likely to use vocabularies familiar to any individual searcher. The best answer is twofold: First, it is desirable to have an index (a “mapping”) from the natural language of each group of searchers to the entries used in each metadata vocabulary. Such a mapping provides an index from a vocabulary familiar to the searcher to the vocabulary used in entries of the target system and so is called a search term recommender system. (We called it an “entry vocabulary index.”) Dewey’s “relativ index” to his Decimal Classification is a familiar example. When searching across databases one also wants a second kind of mapping: between pairs of system vocabularies. Unfortunately, mappings between different vocabularies are rare, expensive, time-consuming, and hard to maintain. (The Unified Medical Language System is a notable example).⁵ Our impression is that this problem is worse in searching across different media forms because data bases in different media forms tend

to be created by different communities, increasing the chances that they will use different categories, vocabularies, and ways of thinking.

Fortunately where data containing two forms of vocabulary are available they can be used as training sets for statistical association techniques to generate entry vocabulary indexes automatically and we used this approach as outlined below. (More details can be found in the Appendix).

From Text words to Library Catalog Subject Headings

An entry vocabulary index from ordinary English words to *Library of Congress Subject Headings (LCSH)* was created by taking catalog records containing at least one subject heading (6xx field in the MARC bibliographic format). From each of the 4,246,510 records used, we extracted main subject headings (subfield a from fields 600, 610, 611, 630, 650 and 651) and fields containing text: titles (245a), subtitles (245b), and summaries describing the scope and general content of the material (520a). The underlying assumption is that for each record the words in our “text” fields (245a,b and 520a) tend to be characteristic of discourse on the subject (6xxa). Two examples, with identifying LCCN numbers in the <001> field are:

```
<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>
```

```
<001>73180255 </001>
<245><a>Reptilian disease</a><b>recognition and treatment</b></245>
<650><a>Reptiles</a><x>Diseases</x></650>
```

The words in the “text” fields (245a, 245b and 520a) were extracted. Stop words were removed and the remainder normalized. Then the degree to which each word is associated with each subject heading (by co-occurring in the same records) was computed using a maximum likelihood ratio-based measure. Natural language processing can be used to identify adjective-noun phrases to support more precise searching using phrases as well as individual words. A very large matrix shows the associatedness of each text word (or phrase) with each subject heading, so, for any given word (or combination of words), a list of the most closely associated headings, ranked by degree of associatedness, can be derived from the matrix.

Queries

A query, which can be a single word, a phrase, a set of keywords, a book title, and so on is normalized in the same way and looked up in the matrix to produce a ranked list of the most closely associated subject headings as candidate LCSH search terms. For example, entering the textual query words “Peanut” and “Butter” generates the following ranking list of LCSH main headings as candidates for searching:

Rank	LCSH (subfield 650a)
1	Peanut
2	Cookery (peanut butter)
3	Cookery (peanuts)
4	Peanut industry
5	Peanut butter
6	Butter
7	Schulz, Charles M.

This display is an important departure from traditional fully automatic searching. The list is, in effect, a prompt, indicating probably suitable query terms in the vocabulary of the target resource. It introduces the searcher to the categories and terminology of the system and enables the searcher to use expert judgment to select the heading that seems best for the search.

From Text words to the Metadata Vocabularies in Numeric Data Sets

A training set of records containing both descriptive words and topical metadata is often not readily available for numeric data sets. Our first effort was to create an Entry Vocabulary Index to the Standard Industrial Classification (SIC), widely used over many years in numeric data sets. (We associated SIC codes with words by using, as a training set, the titles in a bibliographic database which used SIC codes.) But by the time our SIC Entry Vocabulary Index was completed, the SIC had been discontinued and replaced by the North American Industry Classification System (NAICS), so we created a mapping from SIC codes to NAICS codes. Figures 1-3 show stages in an interface that accepts a searcher's query "car" (Figure 1), prompts with a ranked list of NAICS codes (Figure 2), then extends the search with the selected NAICS code to retrieve numeric data (Figure 3).

Address <http://otlet.sims.berkeley.edu/geonaics.html>

Links [VIPE](#) [INEX](#) [AGW](#) [Amazon](#) [Going places](#) [Google](#) [hotmail](#) [InfoAccess](#)

Entry Vocabulary Dictionary for NAICS (North American Industry Classification System)

Enter keywords to search the Entry Vocabulary Dictionary of NAICS classification codes. A list of the NAICS codes most closely associated with the terms in your request will be returned.

Enter a keyword to search here:

Figure 1: Query interface for search term recommender system for the North American Industry Classification System.

Search Results for keyword "car"

Select a NAICS code and a database and press the "Search" button to retrieve the data from the US Census site

Choose a level of geography:

United States Total
 State Total
 County Total

United States ▼ If you selected "State" or "County" level, please select a state

Search 1997 Economic Census database
 Search 1998 County Business Patterns

database Search

SELECT	ID	Rank	NAICS CODE	Description
<input type="radio"/>	49049	119.15	336111	Automobile Manufacturing
<input type="radio"/>	49578	68.40	336322	Other Motor Vehicle Electrical and Electronic Equipment Manufacturing
<input type="radio"/>	102857	61.18	3363	Motor Vehicle Parts Manufacturing
<input type="radio"/>	51233	43.96	336999	All Other Transportation Equipment Manufacturing

Figure 2: Display of NAICS code search term recommendations for “car.”

Address http://www.census.gov/epcd/ec97/mj/M1000_31.HTM#N336

Links [WPe](#) [INEX](#) [AGW](#) [Amazon](#) [Going places](#) [Google](#) [hotmail](#) [InfoAccess](#)

NAICS code	Description	Estab-lish-ments	Value of Shipments (\$1,000)	Annual payroll (\$1,000)	Paid employees
336	Transportation equipment mfg	1,100	108,942,261	13,496,505	268,015
3361	Motor vehicle mfg	35	55,928,384	3,686,803	61,552
33611	Automobile & light duty motor vehicle mfg	31	55,884,009	3,682,799	61,431
336111	Automobile mfg	22	34,665,392	2,517,653	41,470
336112	Light truck & utility vehicle mfg	9	21,218,617	1,165,146	19,961
33612	Heavy duty truck mfg	4	44,375	4,004	121
3362	Motor vehicle body & trailer mfg	78	585,133	109,216	3,583
336211	Motor vehicle body mfg	32	224,653	48,817	1,488

Figure 3: Display of numeric data retrieved using selected NAICS code.

By this time, however, it had become apparent that, with the current low level of interoperability in software and in data formats, the labor required to create Entry Vocabulary Indexes and interfaces to each large traditional numeric data series was so large that we turned our attention to a collection of different numeric data sets available through a single interface, *Counting California*, made available by California Digital Library at <http://countingcalifornia.cdlib.org/>. This resource 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, the California Department of Finance, and the federal Bureau of the Census. The tables are organized under a two-level classification scheme. There are 16 topics at the top level, which are subdivided into a total of 184 subtopics. All the numeric tables assigned to one or more subtopics and each table has a caption.

At the Counting California website a searcher can browse for tables by selecting a higher level topic, then a lower level subtopic, and then a table. We created two additional ways to access the tables: Probabilistic retrieval, and an Entry Vocabulary Index to the topical categories. We extracted the captions, topics, and subtopics for each of the 3,000 tables and created XML records in the following form:

```
<table>
<topic> education </topic>
<subtopic> libraries </subtopic>
<caption> library statistics, statewide summary by type of library California 1992-93 to 1997-98
</caption>
</table>
```

Retrieval

We used two search methods:

I. *Direct Probabilistic Retrieval*. We used an in-house implementation of a probabilistic full-text retrieval algorithm developed at Berkeley.⁶ This search engine takes a free form text query and returns a ranked list of captions of tables ranked according to their relevance scores. For example, the five top-ranked captions returned to the query “Public Libraries in California” were:

1. Library statistics, Statewide summary by type of library California, 1992-93 to 1997-98 Table F6.
2. Library statistics, Statewide summary by type of library California, 1993-94 to 1998-99 Table F6YR0-0.
3. Number of California libraries, 1989 to 1999 Table F5YR00
4. Number of California libraries, 1989 to 1998, as of September Table F5.
5. California Public Schools, Grades K-12, 1989 to 1998 Table F4.

Each entry in the retrieved set list is linked to a numeric table maintained at the Counting California website and, by clicking on the appropriate link, a user can display the table as an MS Excel file or as a pdf file.

II. *Mediated Search*. From the same extracted records we used the words in the captions to create an Entry Vocabulary Index to the subtopics in the topic classification using the method already described. As an example, the query “personal individual income tax,” when submitted to the Entry Vocabulary Index, generated the following ranked list of subtopics:

1. Income
2. Government earnings and tax revenues
3. Personal income
4. Property tax
5. Personal income tax
6. Corporate income tax
7. Per capita income

A user can click on any selected subtopic to retrieve the captions of tables assigned that subtopic. For example, clicking on the fifth subtopic Personal income tax retrieves:

- Personal income tax returns: Number and amount of adjusted gross income reported by adjusted gross income class California, 1998 taxable year. Table D10YR00
- Personal income tax returns: Number and amount of adjusted gross income reported by adjusted gross income class California, 1997 taxable year. Table D9
- Personal income statistics by county, California 1997 taxable year. Table D10
- Personal income statistics by county, California 1998 taxable year. Table D11YR00

TRANSVERSE SEARCHING BETWEEN TEXT AND NUMERIC DATA SERIES

To demonstrate the searching capability from a bibliographic record to numeric data sets, 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.

Searching from Catalog Records to Numeric Data Sets

Having selected some text, in our case a catalog record, how could one identify the facts or statistics in a numeric database that are most closely related to the topic? Clicking on a “formulate query” button placed at the end of a displayed full MARC record creates a query for searching a numeric database. The initial query will contain the words extracted from the title, subtitle, and the subject headings and is placed in a new window where the user can modify or expand the query before submitting it to the search engine for a numeric database. So, for example, the following text extracted from a catalog record:

Library laws of the State of California,
Library legislation. California.
Public libraries

when submitted as a query, retrieves a ranked list of table names, of which two, covering different time periods, are entitled *Library Statistics, Statewide Summary by Type of Library, California*.

Searching from Numeric Data Sets from Catalog Records

Transverse search in the other direction, starting from a data table, is achieved by forwarding the caption of a table to the word-to-*LCSH* entry vocabulary index to generate a prompt list of the seven top-ranked *LCSH* headings, any one of which can be used as a query submitted to the catalog.

ARCHITECTURE

Figure 4 shows the structure of the implementation. The boxes shown in the figure are:

1. A search interface 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.
4. An online catalog.
5. Results of searching the online catalog using a *LCSH*.
6. A full MARC record displayed in tagged form.
7. A new query formed by extracting the title and subject fields from the displayed full MARC record.
8. A numeric database.
9. A list of captions of numeric tables ranked by relevance score to the query.
10. Numeric table displayed in PDF or MS Excel format.
11. A search interface for numeric databases based on a probabilistic search algorithm.

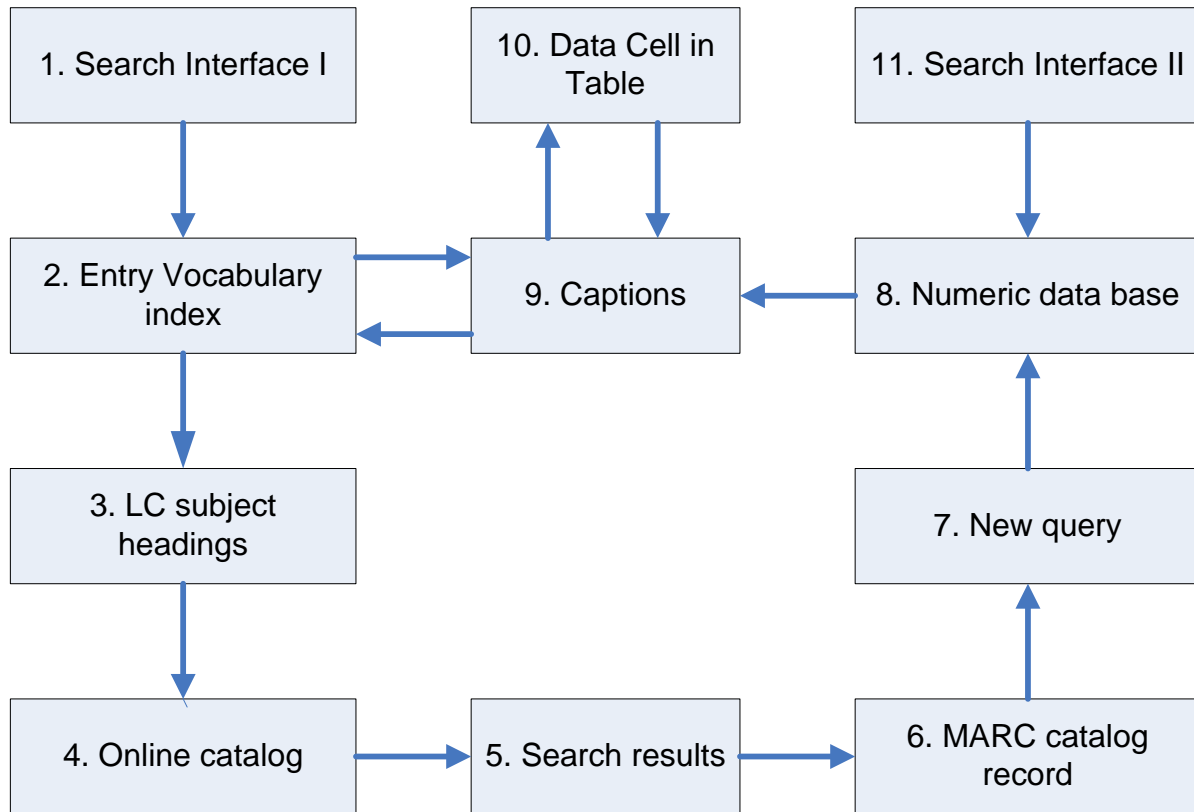


Figure 4: Architecture of the prototype.

A user can start a search using either interface (boxes 1 or 11), and, from either starting point, find records on the same topic of interest in a textual (here bibliographic) database and a socio-economic database.

CONCLUSIONS AND FURTHER WORK

1. Enhanced Access to Numeric Data Sets

The descriptive texts associated with numeric tables, such as the caption, headers, or row labels, are usually very short. They provide a rather limited basis for locating the table in response to queries, or describing a data cell sufficiently to form a usefully descriptive query from it. Sometimes the title (caption) of a table may be the only searchable textual description about the content of the table and the titles are sometimes very general. For example, one of the titles noted above, *Library Statistics, Statewide Summary by Type of Library California, 1992-93 to 1997-98*, is so general that neither the kinds of statistics nor the types of libraries are revealed. If a user posed 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 words in common between the table title and the user’s query is ‘California’ and, if the plurals of nouns have been normalized, to the singular form, ‘library.’

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. We recommend that the textual descriptions of numeric tables be enriched. Automatically combining the table title and its column and row headings would be a small but practical step towards improved retrieval.

2. Geographic Search

Socio-economic numeric data series refer to particular areas and, in contrast to text searching, the geographical aspect ordinarily has to be specified. To match the geographical area of the numeric date, a matching text search may also have to specify the same 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. The problem is compounded because, in numeric data series, specialized geo-political divisions, such as census tracts and counties, are commonly used. These divisions do not match conveniently with searchers’ ordinary use of place-names. Also, the granularity of geographical coverage may not match well. Data relating to Berkeley, for example, may be available only in aggregated data for Alameda County.

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. This means that gazetteers and map visualizations become important. Gazetteers relate named places to defined spaces, and thereby reveal spatial relationships between places, e.g. the city of Alameda is on Alameda island within Alameda County. This problem that has been addressed in a subsequent study entitled “Going Places in the Catalog: Improved Geographical Access.”⁷

3. Temporal Search

Searches of text files and of socio-economic numeric data series also differ substantially with respect to time periods: Numeric data searches ordinarily require the years of interest to be specified; text searches rarely specify the period. An additional difficulty arises because in text, as in speech, a period is commonly referred to by a name derived metaphorically from events used as temporal markers, rather than by calendar time, as in “during Vietnam,” “under Clinton,” or “in the reign of Henry VIII.”

Named time periods have some of the characteristics of place names: they are culturally based and tend to be multiple, unstable, and ambiguous. It appears that an analogous solution is indicated: directories of named time periods mapped to calendar definitions, much as a gazetteer links place names to spatial locators. This problem is being addressed in a subsequent study entitled “Support for the Learner: What, Where, When, and Who.”⁸

4. Media Forms

The paradox, in an environment of digital “media convergence,” that it appears

impossible to search directly across different media forms invites closer attention to concepts and terminology associated with media. A view that fits and explains the phenomena as we understand them, distinguishes three aspects of media:

(a) *Cultural codes*: All forms of expression depend on some shared understandings, on language in a broad sense. Convergence here means cultural convergence or interpretation.

(b) *Media types*: Different types of expression have evolved: Texts, images, numbers, diagrams, art. An initial classification can well start with the five senses of sight, smell, hearing, taste, and feel.

(c) *Physical media*: Paper; film; analog magnetic tape; bits; . . . Being digital affects directly only this aspect.

Anything perceived as a meaningful *document* has cultural, type, and physical aspects and *genre* usefully denotes specific combinations of code, type, and physical medium adopted by social convention. Genres are historically and culturally situated.

Convergence can be understood in terms of interoperability and is clearly seen in physical media technology (c, above). The adoption of English as a language for international use in an increasingly global community promotes convergence in cultural codes (a). Nevertheless, the different media types (b) are fundamentally distinct.

5. Metadata as infrastructure

It is the metadata and, in very broad sense, “bibliographic” tools that provide the infrastructure necessary for searches across and between different media: Thesauri, mappings between vocabularies, place name gazetteers, and the like. In isolation, metadata is properly regarded as description attached to documents, but this is too narrow a view. *Collectively*, the metadata forms the infrastructure through which different documents can be related to each other. It is a variation on the role of citations: Individually references amplify an individual document by validating statements made within it; collectively, as a citation index, references show the structure of scholarship to which documents are attached.

SUMMARY

A project was undertaken to demonstrate simultaneous search of two different media types (socio-economic numeric data series and text files) without ingesting these diverse resources into a shared environment. The project objective was eventually achieved, but proved harder than expected for the following reasons: Access to these different media types have been developed by different communities with different practices; the systems (vocabularies) for topical categorization vary greatly and need interpretative mappings (also known as relative indexes, search term recommender systems, and entry vocabulary indexes); specification of geographical area and time period are as necessary for search in socio-economic data series and, for this, existing procedures for searching text files are inadequate.

APPENDIX: STATISTICAL ASSOCIATION METHODOLOGY

A statistical maximum likelihood ratio weighting technique⁹ was used to construct a two-way contingency table relating each natural language term (word or phrase) with each value in the metadata vocabulary of a resource, e.g. Library of Congress Subject Headings (LCSH), Library of Congress Classification Numbers, U.S. Patent Classification Numbers, and so on. An associative dictionary that will map words in natural languages into metadata terms can also, in reverse, return words in natural language that are closely associated with a metadata value.

Training records containing two different metadata vocabularies can be used to create direct mappings between the values of the two metadata vocabularies. For example, U.S. patents contain both U.S. and International Patent Classification numbers and so can be used to create a mapping between these two quite different classifications. Multilingual training sets, such as catalog records for multilingual library collections, can be used to create multilingual natural language indexes to metadata vocabularies and, also, mappings between natural language vocabularies.

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 can be used in creating association dictionaries.

The training set used to create the word-to-LCSH entry vocabulary index was a set of catalog records with at least one assigned Library of Congress Subject Heading (i.e., at least one 6xx field). Natural language terms were extracted from the title (field 245a), subtitle (245b), and summary note (520a). These terms were 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 considered not to be content bearing, such as pronouns, prepositions, coordinators, determiners, and the like.

The content words (those not treated as stopwords) are normalized using a table derived from an English morphological analyzer.¹⁰ The table maps plural nouns into singular ones; verbs into 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 past tense of the verb *to see* in, is not reduced to *see*. Subject headings (field 6xxa) were extracted without qualifying subdivisions. The inclusion of foreign words (*alcoholismo*, *alcoholisme*, *alcohol*, and *alcool*), derived from titles in foreign languages, demonstrate that the technique is language-independent and could be adopted in any country. It could also support diversity in U.S. libraries by allowing searches in Spanish or any other languages, so long as the training set contains sufficient content words. Entry vocabulary indexes are accessible at <http://metadata.sims.berkeley.edu/prototypesI.html>.

Fuller description of the project methodology can be found in ¹¹.

REFERENCES

¹ Michael K. Buckland, Fredric C. Gey, and Ray R. Larson, *Seamless Searching of Numeric and Textual Resources: Final Report on Institute of Museum and Library Services National Leadership Grant No. 178*. (Berkeley, CA: University of California, School of Information Management and Systems, 2002),

<http://metadata.sims.berkeley.edu/papers/SeamlessSearchFinalReport.pdf> (accessed July 18, 2006); and Michael Buckland, Aitao Chen, Fredric C. Gey, and Ray R. Larson, "Seamless Searching of Numeric and Textual Resources: Friday Afternoon Seminar, Feb. 14, 2003," 2003, <http://metadata.sims.berkeley.edu/papers/seamlessfri.ppt> (accessed July 18, 2006).

² Michael Buckland et al., "Mapping Entry Vocabulary to Unfamiliar Metadata Vocabularies," *D-Lib Magazine* 5, no. 1 (January 1999), <http://www.dlib.org/dlib/january99/buckland/01buckland.html> (accessed July 18, 2006); Michael Buckland, "The Significance of Vocabulary," 2000, <http://metadata.sims.berkeley.edu/vocabsig.ppt> (accessed July 18, 2006); and 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* (San Francisco: Morgan Kaufmann, 2001), 91-95, <http://metadata.sims.berkeley.edu/papers/hlt01-final.pdf> (accessed July 18, 2006).

³ Michael Buckland, "Vocabulary as a Central Concept in Library and Information Science," in *Digital Libraries: Interdisciplinary Concepts, Challenges, and Opportunities. Proceedings of the Third International Conference on Conceptions of Library and Information Science (CoLIS3), Dubrovnik, Croatia, 23-26 May 1999*, ed. T. Arpanac et al. (Lokve, Croatia: Benja Publishing, 1999), 3-12, <http://www.sims.berkeley.edu/~buckland/colisvoc.htm> (accessed July 18, 2006); and Buckland et al., "Mapping Entry Vocabulary."

⁴ *Counting California*, <http://countingcalifornia.cdlib.org> (accessed July 18, 2006).

⁵ "Factsheet: Unified Medical Language System," www.nlm.nih.gov/pubs/factsheets/umls.html (accessed July 18, 2006).

⁶ 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, ed., *The Second Text REtrieval Conference (TREC-2), March 1994*, 57–66 (Gaithersburg, MD: National Institute of Standards and Technology, 1994), <http://trec.nist.gov/pubs/trec2/papers/txt/05.txt> (accessed July 18, 2006).

⁷ "Going Places in the Catalog: Improved Geographical Access," <http://ecai.org/imls2002> (accessed July 18, 2006).

⁸ Vivien Petras, Ray Larson, and Michael Buckland, "Time Period Directories: A Metadata Infrastructure for Placing Events in Temporal and Geographic Context," in *Opening Information*

Horizons: Joint Conference on Digital Libraries (JC DL), Chapel Hill, NC, June 11-15, 2006, forthcoming, <http://metadata.sims.berkeley.edu/tpdJC DL06.pdf> (accessed July 18, 2006); “Support for the Learner: What, Where, When, and Who,” <http://ecai.org/imls2004> (accessed July 18, 2006).

⁹ Ted Dunning, “Accurate Methods for the Statistics of Surprise and Coincidence,” *Computational Linguistics* 19 (March 1993), 61–74.

¹⁰ Karp, Daniel, Yves Schabes, Martin Zaidel, and Dania Egedi, “A Freely Available Wide Coverage Morphological Analyzer for English,” in *Proceedings of COLING-92, Nantes, 1992*, (Morristown, NJ: Association for Computational Linguistics, 1992), 950-955, <http://acl.ldc.upenn.edu/C/C92/C92-3145.pdf> (accessed July 18, 2006).

¹¹ Buckland, Gey, and Larson, *Seamless Searching*; Youngin Kim, Barbara Norgard, Aitao Chen, and 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 29-Nov 4, 1999* (Medford, NJ: Information Today, 1999), 172-180.