

Electronic Cultural Atlas Initiative, University of California, Berkeley

Going Places in the Catalog: Improved Geographic Access

Final Report

A project supported by the
Institute of Museum and Library Services
National Leadership Grant LG-02-02-0035-02, Oct 2002 - Sept 2004.

Principal Investigators: Michael K. Buckland, Fredric C. Gey, and Ray R. Larson

INTRODUCTION

Libraries and museums have a broad need to support the searching by place. This can be done directly (e.g. FIND **Andorra**) or indirectly through geographical aspects of any topic (e.g. FIND **Folklore – Andorra**). Either way, geographic aspects need to be represented in the description of objects.

This project originated in an earlier IMLS-supported project entitled *Seamless Searching of Numeric and Textual Resources* which was concerned with extending topical searches to and between two different genres: textual databases and socio-economic data series. This objective was achieved, but it became evident that topical searching in socio-economic data series ordinarily requires that the geographical area also be specified, that data series tend to specify geographic area in specialized ways that differed from text databases, that place names alone were too varied to for reliable searching, that spatial definitions (latitude and longitude) were needed, and that map displays were very important for the searchers' comprehension of what was being done (See <http://metadata.sims.berkeley.edu/GrantSupported/seamless.html>)

Places and Spaces

There is a basic distinction between *place*, a cultural concept, and *space*, a physical concept. Cultural discourse tends to be about *places* rather than *spaces* and, being cultural and linguistic, *place names* tend to be multiple, ambiguous, and unstable. Indeed, the places themselves are unstable. Notoriously, cities expand, absorbing neighboring places, and countries change both names and boundaries. Spaces are more stable and are ordinarily specified by physical measurements. Each place can be characterized by the space it is in and so two or more places can be related to each other by reference to a common system of spatial description (“georeferencing”). The internationally accepted system of latitude and longitude provides a common standard and, very importantly, georeferencing allow both location and spatial relationships to be visualized in map displays.

Places in Library Catalogs

Support for geographical searching in library catalogs is provided primarily by using place names as geographic subject headings (e.g. MARC 651) or as geographic aspects (“geographic subdivisions”) of other kinds of subject headings (e.g. MARC 65X \$z) qualifying other kinds of subject headings, usually topics (e.g. 650\$a**Folklore**\$z**Andorra**). This approach is weak for two reasons: First, *place names* are used; and, second, the places named are typically

political jurisdictions, which are themselves unstable. (Poland ceased to be a country from 1795 through 1918. Even if you knew that, how would you search for that area during that time?)

In this project we explored two quite different kinds of remedy: First, making better use of existing geographical clues in catalog records; and, second, using gazetteers and map displays to replace or augment place name authority records.

Two caveats. As will be shown, the data in library catalogs reflect descriptive cataloging rules of great complexity. Many descriptive codes are rarely encountered. They may well be valid yet are unlikely to be used in searches because unexpected (e.g. the Geographic Area Code *c* for *Cold Regions*). Further, the rules are continuously evolving, so research libraries' catalogs contain large quantities of legacy records, some created in the nineteenth century, which have not been revised as the rules change. For both reasons, a study of this kind requires some simplification for concise, coherent conclusions.

SAMPLE CATALOG RECORDS

Geographic clues are in **bold** and our notes in *italic* in these abbreviated MARC records.

Isle of Man Tramways. ISBN 0715347403

- 008 700812 1970 **enkabh**, b,fe 001 0 eng *Country of publication code for England*
- 043 a **e-uik**– *Geographic Area Code. The cataloger has erroneously used the “Country Code” used in field 008, instead of the Geographic Area Code prescribed for 043. Should be e-uk-ui for “Europe. Great Britain Miscellaneous Island Dependencies”.*
- 050 0 a TF764.M27 b P4 1970 *Geographic code within a Library of Congress Classification number*
- 082 0 a 388.4/6/09**4289** *Geographic code within a Dewey Decimal Classification number. An error. Should be 4279.*
- 100 1 a Pearson, Frederick Keith. *Author*
- 245 10 a **Isle of Man** tramways, c by F. K. Pearson;... *Place name used adjectivally in title.*
- 260 a **Newton Abbot** : b David & Charles, c 1970. *Place of publication, not in the Isle of Man.*
- 500 a Imprint covered by label: A. M. Kelley, **New York**. *Note that Place of publication obscured.*
- 6102 0 a **Manx** Electric Railway Company. *Adjective for [Isle of] Man in a corporate name used as subject heading.*
- 650 0 a Street-railroads z **Man, Isle of**. *Geographic subdivision using inverted form of name. In this record the island known as Man is represented six different ways.*

History of Turner County

- 008 941107r19931933**miuabc** b 000 0deng d *Microfiche created in Michigan, USA. Original, published in Georgia, USA would have been gau*
- 043 a **n-us-ga** *Geographic Area Code*
- 050 4 a **F292.T8** b P2 *Library of Congress Classification number for Turner County, NY.*
- 100 1 a Pate, John Ben, d 1874- *Author*
- 24510 a History of **Turner County** h [microform] / c by John Ben Pate. *Place name in Title.*
- 260 a **Atlanta, Ga.** : b Stein Print. Co., c 1933. *Place of publication of original*

- 651 0 a **Turner County (Ga.)** x History. *Place name, with state abbreviation a “qualifier,” used as a subject heading.*
- 651 0 a **Turner County (Ga.)** x Genealogy. *Place name as subject heading.*
- 650 0 a Registers of births, etc. z **Georgia z Turner County** *Geographical subdivisions, using a hierarchical structure instead of a qualifier.*

New Amsterdam and its People

- 008 710331r19691902nyuace 000 0 eng *Place of publication code for New York state.*
- 043 __ a **n-us-ny** *Geographic Area Code for New York state.*
- 050 00 a **F128.4** |b .I58 1969 *Geographic code for New York city as part of a Library of Congress Classification number.*
- 082 00 a **974.71/02** *Dewey Decimal Congress Classification number for New York city.*
- 100 1_ a Innes, J. H. |q (John H.) *Author.*
- 245 10 a **New Amsterdam** and its people; |b studies, social and topographical, of the town under Dutch and early English rule, ... *Obsolete and ambiguous place name.*
- 260 __ a **Port Washington, N.Y.**, |b I. J. Friedman |c [1969] *Place of publication*
- 440 _0 a **Empire State** historical publications series, |v no. 63 *Vernacular place name in series note.*
- 651 _0 a **New York (N.Y.)** |x History |y Colonial period, ca. 1600-1775. *Place name with state as qualifier.*

These records illustrate only the commoner examples. There are several other options, including 044 Country of publishing/producing entity code, 052 Geographic scope expressed using a classification scheme, and note fields, including 545 Biographical or historical data.

CATALOG RECORD ENRICHMENT

Library catalog records contain more geographical clues than are used and our first objective was to examine whether and how better use could be made of a wider range of data within the record. We reached three conclusions on three aspects:

1. *Geographical Area Code (MARC 043)*. The MARC Bibliographic format includes field 043, a Geographical Area Code system. The most commonly found codes are in the general form [Continent] – [Country] – [Province or state (sometimes)], e.g. *e-lu* for Luxemburg and *n-us-id* for Idaho. The coding system is, however, much more complex. The Indian Ocean has hyphenated country divisions, , like the six continents, but the Atlantic and Pacific oceans do not. There is also provision for regions (e.g. *aw Middle East*), border regions (*m Intercontinental areas (Eastern Hemisphere)*), classes of countries (e.g. *d Developing countries*), geographical features (e.g. *fr Great Rift Valley*), political groups (e.g. *b Commonwealth countries*), climatic zones (e.g. *q Cold regions*), and extraterrestrial places (e.g. *zd Deep Space*). The great majority appear to be rarely assigned. We know of no library catalog that allows search on this code, so adding that capability is an obviously feasible enhancement.

The significance of the 043 Geographical Area Code becomes much greater if, as we had expected, many records had 043 codes but not geographical codes in the Subject Headings (651 or 65X \$z), or vice versa. It should be feasible to add enrich the record algorithmically by generating 043 codes from the geographical codes in the Subject Headings. Likewise, 043 codes

could be used to generate geographical codes to enrich the Subject Headings, which would be more useful since Subject Headings are searchable in current online catalogs.

A set of MARC records from the MELVYL union catalog of the University of California libraries was kindly made available to us for our research by the California Digital Library. This allowed us to perform a statistical analysis on a set of five million MARC records, all originating from the Library of Congress, but with call numbers and possibly other modifications made by University of California libraries. We found that, to a greater extent than expected, that records with geographical codes in the Subject Headings did in fact also have 043 Geographic Area Codes and vice versa.

Less than 4% of the records with one or more geographic subject headings (650z and/or 651) did not also have a 043 Geographic area code and less than 5% of records with a 043 Geographic area code did not have a geographic subject heading. The simple conclusion is that there is little scope for catalog record enrichment by inferring missing 043 Geographical Area codes from Geographical Subject Headings (651) or Geographical Subdivisions (65X4z). It is a welcome conclusion in that the situation is better than expected. For details see Petras 2004.

2. A similar analysis was performed on language codes on the hypothesis that a book about, say, folklore published in Croatian would probably tend to be about Croatian folklore. Here the hypothesis was found to be valid, but the conclusion needs to be stated carefully. Based on the University of California catalog records analyzed, English language books may be on any topic and any place, but foreign language books tend to be about the place(s) in which the language is used. This is, we assume, generally true in the context of U.S. academic library collection development policies. The implication is that in collection development, books in English are preferred and then, to strengthen holdings about specific foreign countries, titles published in that country (or region) are added and these tend to be in the language of that country.

The same is true, within library collections, to books *published* in foreign countries. The implication is that use of language and/or place of publication can serve as an imperfect but serviceable indicator of geographic scope.

Both conclusions may be true of publishing practices generally, but our evidence relates specifically to library collections, which is what catalogs represent. (Details in Petras 2004)

3. Geographical aspects are encoded within classification numbers, as noted in the sample records above. Using these codes would depend on the geographical fragment being identifiable. This would be straightforward with the Universal Decimal Classification, but that system is very rarely used. With the Dewey Decimal Classification and Library of Congress Classifications, the two most widely used systems, the geographical component would need to be identified and marked up to be usable. We have not attempted to create algorithms to parse existing records for retroactive analysis and mark-up. The rules could be complex especially for the LC Classification. We conclude that this line of development would be unproductive until classification numbers are internally marked up at source.

GAZETTEERS

Standard library cataloging practice is to normalize vocabulary by selecting preferred terms or names and providing cross-references from non-preferred terms and, also, when names change, to successive preferred names. This *vocabulary control* guides the catalogers' choices

and can, in theory be used by catalog searchers. A well designed online catalog would invoke, explain, and deploy cross-references automatically for place names used as subject headings.

Place names in titles provide an important opportunity when searching by keyword, a very popular technique. But this approach is unreliable because no corrective is provided for the ambiguities arising from the instability and multiplicity of place names. In titles, place names are not identified as being place names and, even if they were, no vocabulary control is provided to disambiguate different places with the same name or to link variant names for the same place. This could be done either by marking up place names during cataloging or by natural language parsing techniques. (One simple rule is that words not in a standard dictionary are probably a place, a person or an institution).

The second major problem is lack of connection between library place name authority practice and the *gazetteers*, the well-established place name authority genre developed in geography. Gazetteers are familiar as the long lists of place names printed in the back of atlases. There is, as yet, no national or international standard for gazetteer content or format, but the better gazetteers include at least the following elements:

- The place name;
- The country or area in which the place is located;
- *Feature type*: a coding for the kind of place, e.g. castle, lake, inhabited place, airport, etc.
- Spatial references: latitude and longitude; and
- References to or from other names for the same place.

When used at the back of an atlas the gazetteer is used as an index to the place names printed on the maps and so each entry also refers to the page(s) or map(s) where the place name is can be found. But the gazetteer is a valid genre in its own right. It is, in effect, for place names what a biographical dictionary is for persons or a business directory for firms. A notable U.S. example is the gazetteer made available by the National Geospatial-Intelligence Agency (NGA) and the Board on Geographic Names (BGN) and often still referred to by its former name the National Imagery and Mapping Agency (NIMA) gazetteer (<http://www.nga.mil/>). It is searchable as the GEOnet Names Server (GNS) (<http://earth-info.nga.mil/gns/html/index.html>). Another well-known example is the gazetteer and gazetteer service of the Alexandria Digital Library.

The special advantage of a gazetteer is the inclusion of spatial coordinates. Latitude and longitude enable named places to be found. They can also be *shown* on a map, in context, using map visualization software. Feature type codes in gazetteers also hold the promise of supporting more precise searching for specific kinds of places. A library catalog could take advantage of these assets if it were to use a gazetteer instead of (or to augment) standard name authority practice. This idea becomes the more feasible with the emergence of protocols for interrogating online gazetteers, notably the Alexandria Digital Library Gazetteer Server Protocol.

The adequacy and suitability of existing gazetteers for library purposes was examined by the Electronic Cultural Atlas Initiative in a prior project entitled *A Multilingual Gazetteer System for Integrating Spatial and Cultural Resources* (Supported by NSF IM/ITR grant 0114019, 2001-02; Principal Investigator: Lewis Lancaster; project Manager Ruth Mostern). Gazetteers have been developed primarily for contemporary geography and for governmental, industrial, and military needs. For the humanities, social sciences, and libraries a national standard for gazetteer content would need to include some capabilities that are not yet ordinarily present: Feature type categories more suited to cultural and historical studies, time codes to indicate *when*

place name was is use, and support for multiple languages and multiple formats. (More at http://ecai.org/projects/gazetteer/nsf_multisys_abstract.html).

In the present network environment there is no justification for maintaining library place name authority files independently of gazetteers. Libraries' place name authority files should become more gazetteer-like and/or link directly to the rich resources of existing gazetteers. Latitude and longitude are especially important for disambiguation and map displays.

THREE DESIGN EXPERIMENTS

Cebuano Interface. Our first map-based catalog interface was for library catalog records for some 700 books about, or published in, the Cebuano region of the Philippines. First, a collection of catalog records was harvested from many online catalogs worldwide using the queries “Cebu” and “Cebuano” and the Z39:50 search and retrieve protocol in the Cheshire retrieval system <http://cheshire.lib.berkeley.edu/>. The library catalog records were formatted into a tab-delimited spreadsheet. A Cheshire script added georeferencing by searching a Cheshire database of the NIMA names for the Philippines (over 70,000 entries) for each place of publication and geographic subject heading or geographic subdivision in the records.

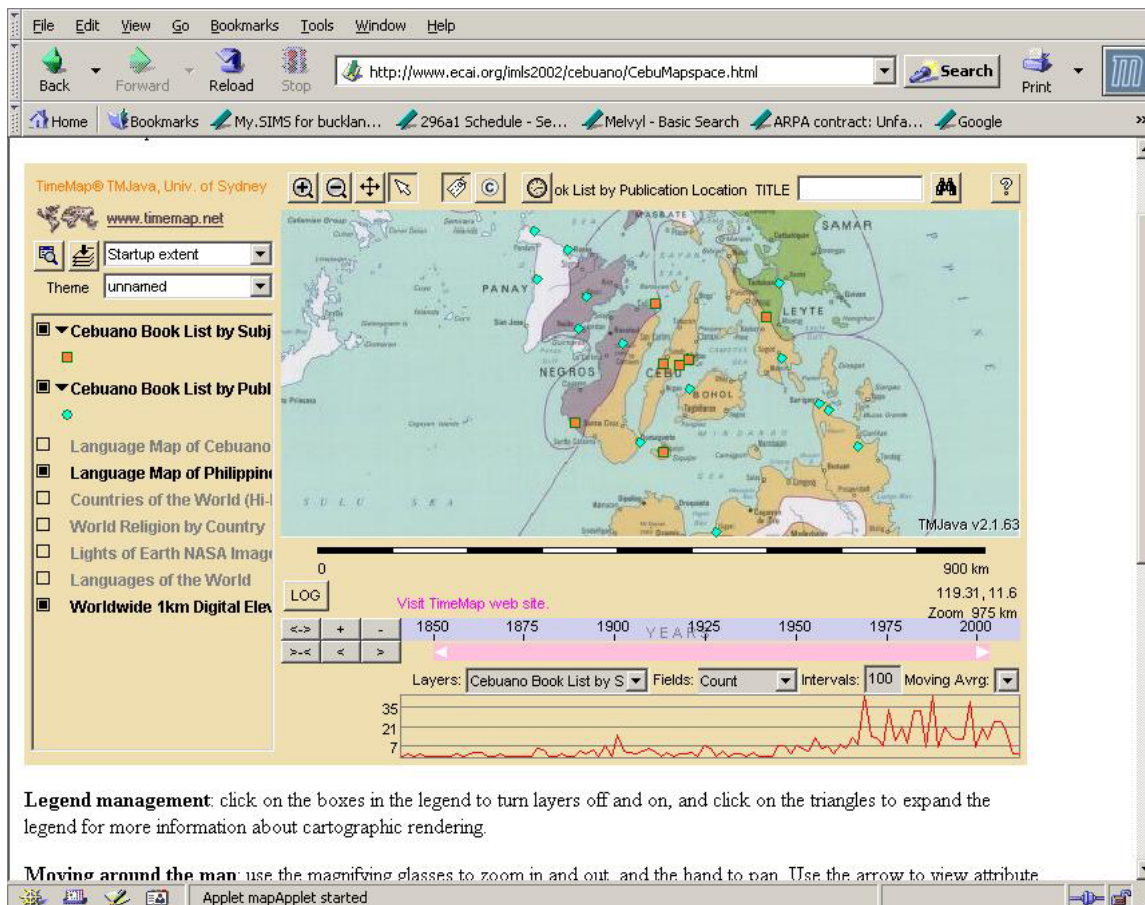


Fig. 1: Geotemporal interface to Cebuano-related catalog records, with language map.

A dynamic map interface supports panning, zooming, and adjusting the time period of interest. (See Fig 1. and note the adjustable time bar below the map.) Small squares marked places contained in subject headings and circles indicated places of publication. Clicking on any square or circle would display the catalog records about, or published at, place respectively

The interface also shows contextual information: the geography of the Cebuano language and other Filipino languages, political boundaries, religious adherence, topography, and other information. This interactive Atlas was created using TMWin tools developed by the TimeMap Project, Archaeological Computing Laboratory, University of Sydney (www.timemap.net). All of the component data sources, with the exception of the scanned language maps (kindly provided by David Blundell, National Taiwan University, and Lawrence Crissman, Griffith University, Australia). To use this interface interactively go to <http://www.ecai.org/imls2002/cebuano/CebuanoIndex.html>

Hindi Language Interface.

The Cebuano interface provided access to catalog, records. Our next interface provided access to actual texts, news reports, which are very much about time and place and so make a good test genre for our purposes. Red squares in Fig 2 show place names extracted from Hindi news documents. The geographic location (latitude and longitude) for each place name was retrieved from a gazetteer of the region, also available for display on the map. A timebar below the map can be adjusted to limit or expand the period to be displayed. The map shows data only for documents whose dates fall within that range. (These images are screen-shots. To use the dynamic TimeMap for Hindi news go to <http://ecai.org/imls2002/hindi/hindimapspace.html>).

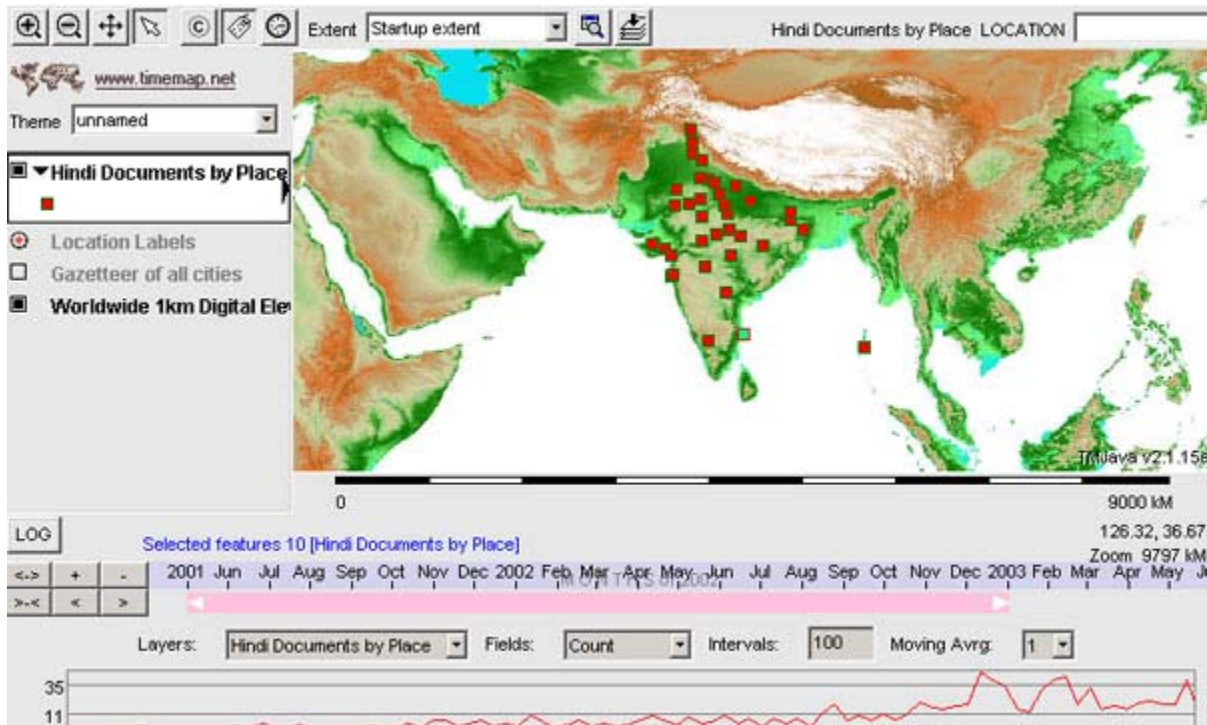
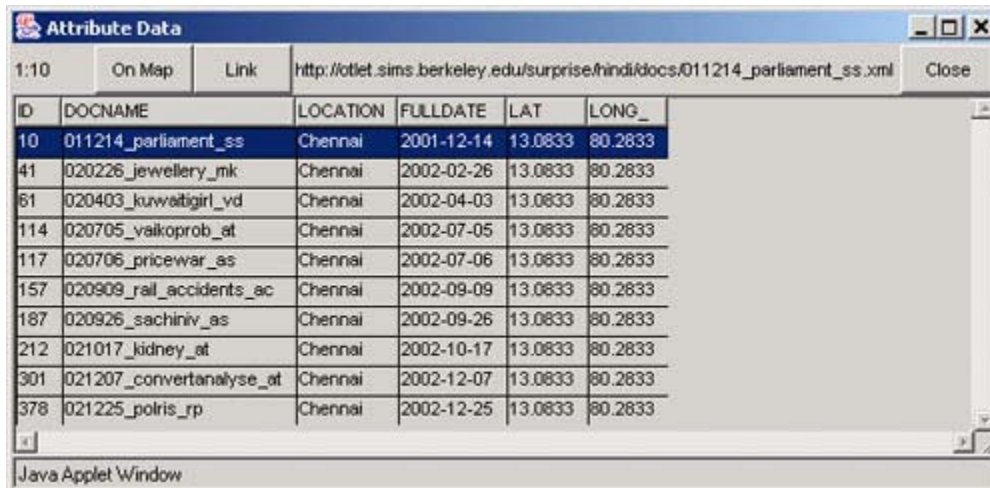


Figure 2: Interface for searching for Hindi news reports by place and time.

Clicking on a single point or by tracing a bounding box around a number of points defines the place or area of interest. The time-bar can be used to limit the temporal range. Brief bibliographic records for the documents that fall within the spatial and temporal limits of your search are displayed in an attribute table (Fig 3).



ID	DOCNAME	LOCATION	FULLDATE	LAT	LONG_
10	011214_parliament_ss	Chennai	2001-12-14	13.0833	80.2833
41	020226_jewellery_mk	Chennai	2002-02-26	13.0833	80.2833
61	020403_kurvaigirl_vd	Chennai	2002-04-03	13.0833	80.2833
114	020705_vaikoprob_at	Chennai	2002-07-05	13.0833	80.2833
117	020706_pricewar_as	Chennai	2002-07-06	13.0833	80.2833
157	020909_rail_accidents_ac	Chennai	2002-09-09	13.0833	80.2833
187	020926_sachiniv_as	Chennai	2002-09-26	13.0833	80.2833
212	021017_kidney_at	Chennai	2002-10-17	13.0833	80.2833
301	021207_convertanalyse_at	Chennai	2002-12-07	13.0833	80.2833
378	021225_polris_rp	Chennai	2002-12-25	13.0833	80.2833

Figure 3: List of news reports found by clicking on Chennai on map.

Selecting one of the records then clicking on the link button will display the text.



Fig 4: Text of news report retrieved by clicking a listed item, then the link button.

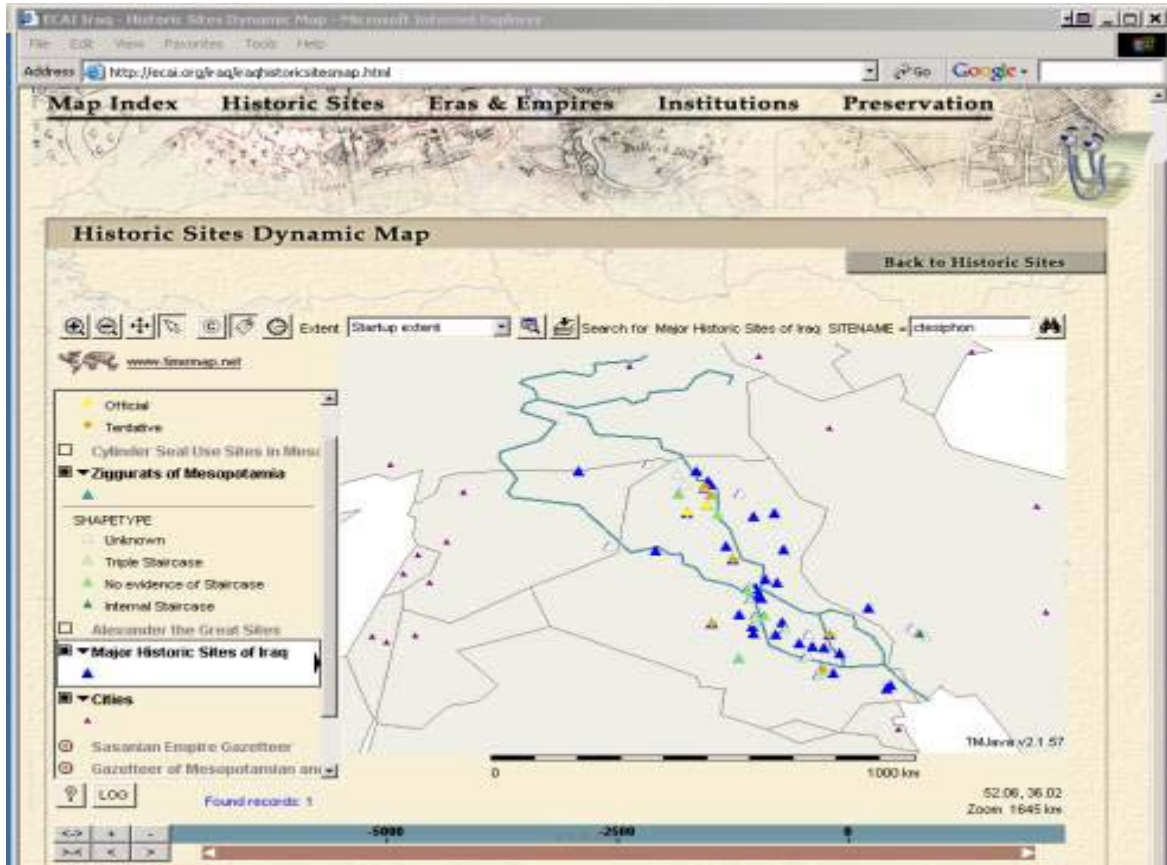


Fig. 5: Map interface in the ECAI Cultural Atlas of Iraq.

Fig 5 shows the map interface of the ECAI Cultural Atlas of Iraq. One can pan and zoom into any desired area and adjust the sliding time bar below to the date range of interest. Only data pertaining to the area shown *and* the specified date range will be displayed.

Clicking the dot for a site will display the underlying data. See Fig. 6 for Ctesiphon.

SITED	SITENAME	ALTERNATEN	FEATURETYP	LATITUDE	LONGITUDE	LOCSSOURCE	STARTDATE	ENDDATE	DATENOTE
7	Ctesiphon			33.0833	44.5833	http://gps.www.rima.nl/gonames/GNSIndex.jsp			

Fig. 6: Gazetteer record for the historic site Ctesiphon

The data for historic site Ctesiphon includes the name of the site, latitude, longitude, feature type, and, among other data, the link to a separate webpage for that site, shown in Fig. 7.

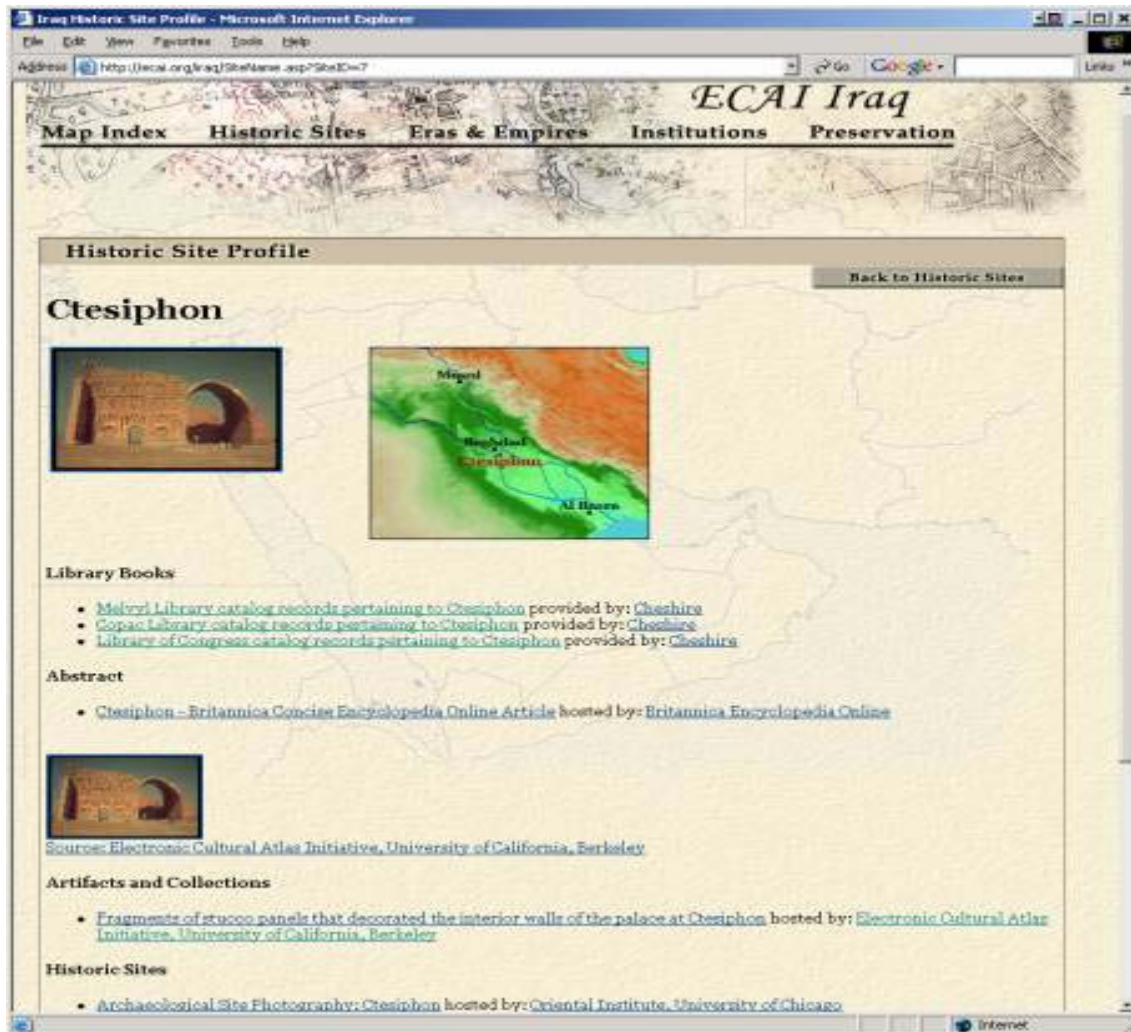


Fig. 7: ECAI Iraq web page for the historic site Ctesiphon.

This page has the customary kinds of links to resources, but note the first section, labeled “Library Books.” These three links generate a search of MELVYL, the union catalog of the hundred libraries of the ten campus University of California; or to COPAC, the union catalog of the largest research university libraries in the United Kingdom; or to the Library of Congress’ online catalog, using the CHESHIRE search and retrieval system. These three catalogs use quite different software, different command languages, and differing search capabilities, but, in each case, the canonical Z39.50 Search and Retrieve protocol is used to interpret and translate both queries and responses to and from the local “languages.” Importantly, these links makes the search as up-to-date as the catalog itself.

Clicking in the third link sends a query about Ctesiphon to the Library of Congress, with the result shown in Fig. 8.



Iraq Historical Atlas
Cheshire II Search Results

Your search, encoded as `zfind ANY ctesiphon`, was submitted to the *Library of Congress* server, **23** items were retrieved.

Record #1

Title:

Arch of Ctesiphon [graphic] : From the N.W. a back view.

Publisher:

[ca. 1933].

Pages:

1 photographic print.

Notes:

Photograph shows view of the Arch of Ctesiphon, Iraq. Use reference copy: DS44.5.M3 1980 P&P Ref. Matson Photo Service Collection (Library of Congress). In album: vol. XI, no. 4659.

Subjects:

[Arches -- Iraq -- Ctesiphon -- 1930-1940. -- lctgm. Photographic prints -- 1930-1940. -- gmjpgc.](#)

LC Call Number:

LOT 11108

URL:

[Click here for document](#)

Fig. 8: First record generated by a search of Library of Congress catalog.

In this instance, there is a “*Click here for document*” link, which, when clicked, leads to the actual document cataloged, a stereo photograph of circa 1933 (Fig. 9) which provides a slightly different perspective on Ctesiphon from the image on the ECAI Iraq website. In this way the ECAI Cultural Atlas of Iraq can provide access to libraries’ documents as well as to catalog records.



Digital ID: cph 3b21221 **Source:** b&w film copy neg.
Reproduction Number: LC-USZ62-73941 (b&w film copy neg.)
Repository: Library of Congress Prints and Photographs Division Washington, D.C. 20540 USA
[Retrieve uncompressed archival TIFF version](#) (1,185 kilobytes)

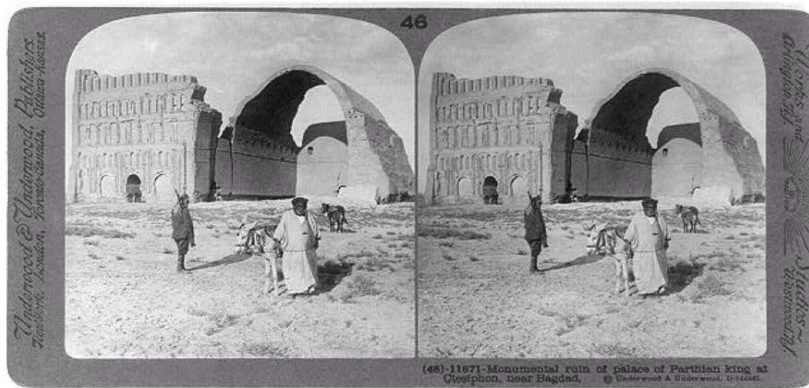


Fig 9: Display of document retrieved from link on catalog record.

THE “GOING PLACES” INTERFACE

To explore the issues raised in this project we built a prototype interface able to:

- Search an online catalog;
- Search an online gazetteer;

Placing the cursor on one of the dots in the map will show the catalog record number(s) in the status bar. Clicking on the dot will also open a new window displaying the records that are associated with that location on the map.

Geographic search from a catalog record

Highlighting any geographic name in the detailed MARC display of a record, then clicking the "yes" check box for the gazetteer search, will search the gazetteer for that name and open a new window displaying place names found and changes the map display to show all the locations with that name. See Fig. 11.

The screenshot shows a software interface with several windows. The main window displays a MARC record with the following text:

```

050 #a BS625
#b .N53 2000
082 #a 221.8/398352
#2 21
100 #a Miditch, Susan.
240 #a Underdogs and tricksters
245 #a A prelude to biblical folklore :
#b underdogs and tricksters /
#c Susan Miditch.
260 #a Urbana, Ill. :
#b University of Illinois Press,
#c 2000.
300 #a xix, 186 p. ;
#c 22 cm.
500 #a Originally published: Underdogs and tricksters
504 #a Includes bibliographical references and indexes
650 #a Folklore in the Bible.
630 #a Bible.
#n o t
  
```

A dialog box titled "Document" is open, asking "Look up the highlighted place name in NIMA gazetteer?" with "Yes" checked. Below it is a search window with a search box, a "Go" button, and a pull-down menu set to "full". A world map shows several red dots in North America. A spreadsheet window titled "NIMA" displays the following data:

ID	RC	UFI	UNI	DD_LAT	DD_L
6976783	1	""	""	30.56028	-94.956
6583947	1	""	""	42.45444	-77.181
5920950	1	""	""	40.11056	-88.207
6639591	1	""	""	40.10833	-83.752
6089314	1	""	""	39.32583	-77.351
5157761	1	-960070	-1412771	7.4166667	-60.716
6412633	1	""	""	46.93444	-96.411
6234687	1	""	""	37.84222	-93.166
5494805	1	""	""	33.15944	-92.445
5964824	1	""	""	37.55806	-95.399
5947412	1	""	""	40.89833	-85.792
5840573	1	""	""	42.22417	-91.874

At the bottom, a list of search results is visible, including "American folklore", "Archaeology and folklore /", "Arte y folklore en Mexican folkways /", "Balkan folk colour language:", and "Bengali folklore collections and studies, 1800-19".

Figure 11: Result of searching the highlighted catalog record word ("Urbana") in gazetteer.

Placing the cursor over a dot in the map will display the NGA record in the status bar.

Search using the map interface and NGA (NIMA) Gazetteer.

A search can also be initiated from the map interface itself or from the gazetteer search box. Fig. 12 illustrates a search for capital cities in South America performed by selecting the feature type "capital city" ("PPLC") in the pull-down menu for NIMA feature types and using the cursor to drawing a bounding box around South America. The results are displayed both in spreadsheet window and by red dots in the map interface.

Searching the catalog from the map interface

To see if there are catalog records available for a geographic place indicated by a dot on the map interface, one right-clicks on the dot to select a catalog option to send off a search

regarding this place (as in the ECAI Cultural Atlas of Iraq pages). The search result regarding this place name will appear as catalog records on the left display, with red dots on the map interface indicating locations of place names associated with those records, as in Figure 10.

As designed, searches in the gazetteer generate place names as queries to be searched in the catalog. Late in the project we began to consider making explicit use of the NGA feature type codes (such as PPLC) for catalog searching, and so, although the project is formally completed, we are continuing the work by mapping between NGA Feature types to corresponding LC Subject Headings in the hope of achieving interoperability. The result will be imperfect because these two vocabularies differ in scale, in emphasis, and mentality. Nevertheless, we hope for a useful outcome.

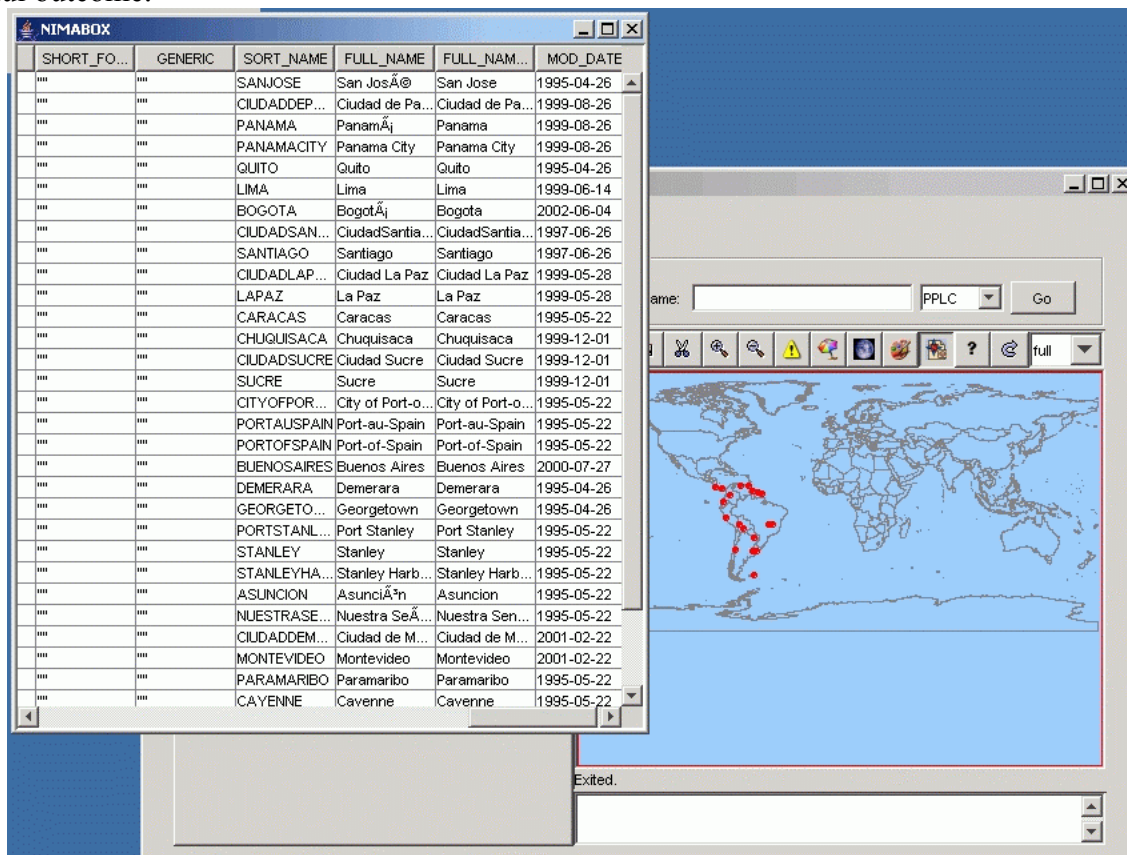


Figure 12. Result of bounding box and feature search in gazetteer.

The interface was designed with three components: The backend containing the local catalog, our copy of the NGA gazetteer, and associated software for using both; software for extending searches elsewhere; and the interface itself. The interface uses Java and so can be used from any web browser without downloading software. However the interface software is extensive and is only suitable when high bandwidth is available. In Fall 2004, once we had had some experience using the Going Places interface, we decided to replace the initial prototype with a new, more sophisticated design, including the NGA feature type to LCSH mapping. This is currently in progress using support from other sources. It will be made available from the Going Places project website when ready for use.

SPATIAL RELATIONSHIPS

Library catalog data typically includes the relationship of containment (one area is entirely within another), which is characteristic of the hierarchical structure of political jurisdictions. MARC 043 *n-us-id* specifies that Idaho is a part of the United States of America which is in the North American continent. Other spatial relationships of interest are *near* (e.g. “within a hundred miles of”), *between*, and *overlapping in area* and these kinds of spatial relationships can be calculated if the latitude and longitude are known.

We implemented *near* and *between* as search arguments using the map display in the interface. The cursor can be used to draw a rectangle (“bounding box”) defining the area between two lines of latitude and two lines of longitude enclosing any one or more points. A square bounding box centered on a point gives a simple, but efficient interpretation of “near”, how near depending on the length of the sides of the box. A rectangular bounding box that just includes two points gives a simple definition of “between”. More complex and precise areas can be used but are computationally very intensive. For example, a bounding circle would provide a more precise expression of “with x miles” of a point. Trigonometry could be used to determine the combinations of latitude and longitude that are within a circle of any given radius of the point of interest. Then the gazetteer would have to be combed for the places that are located within that radius. But both operations require excessive computation. We developed a far less demanding approximation dividing the earth’s surface (and the gazetteer) into many small rectangular tiles, each one with a separate subset of the gazetteer containing only the place names within that tile. A bounding circle is approximated indirectly by determining which tiles are within the circle and then taking all of the places within all of those tiles. Deployment of this tile technique has been deferred into the development of the second version of the interface.

SEARCH PERFORMANCE

Search systems performance can be considered at three levels:

1. *Proof of concept: Does the system actually do what was proposed?* We have described above how the initial prototype interface performed the basic tasks of searching an online catalog, searching an online gazetteer, passing data from the catalog to the gazetteer and vice versa, generating map displays of retrieved sets, and using map displays to aid the formulation of search queries.

2. *Performance evaluation: How well does it perform relative to other systems?*

2.1. Functionality The initial finding is that anything our prototype can find, is probably findable in a second-generation catalog *with enough effort* – if the catalog records are complete, if the data are accurate, and if the searcher brings enough patience and geographical expertise to the search. Library of Congress subject cataloging policy is to assign two different kinds of subject heading for places: A place name with a geographical qualifier and a broader entry for a broader term for that kind of feature, with a geographical subdivision. So, for a particular castle in Austria one assigns the two headings:

- **Schloss Halbturn (Halbturn, Austria)**; and
- **Castles – Austria**.

The combination of a specific instance under its name and the generic category of castles in Austria is important. If you knew of Halbturn castle you could find it if you knew to search

under Schloss Halbturn. A subject keyword search on “Halbturn” ought to include it in a larger retrieved set. It could also be found by tediously scrolling through the result of the broader search on “Castles – Austria”. This system is not precise enough to distinguish castles in an area smaller than Austria, such as Burgenland province where Halbturn is. It would also be inefficient for castles in borderlands. In the area where the Austrian, Italian, and Slovene frontiers meet one would have to use “**Castles -- Austria**”, “**Castles -- Italy**”, “**Castles -- Slovenia**”, and, to be prudent, for older records “**Castles -- Yugoslavia**” and one would still not know without reference to other resources which ones were in that border area. A more extreme case is the model headings for an individual house:

- **2040 Union St. (San Francisco, Calif.)** and
- **Dwellings – California.**

The Going Places interface can support geographic searches more precisely. Drawing a bounding box (or circle) on a map interface allows one to define geographical areas of interest independently of political boundaries. This is especially useful for small areas within large countries, for borderland regions, where country boundaries have changed over time, and when the search covers multiple countries.

2.2. Direct comparison. Ideally, in comparative performance evaluation, two or more systems are used to search the same database. Since our “local catalog” is composed of a subset of records in the University of California MELVYL union catalog, that subset can be used as the common set for comparison. We are well-positioned to compare our prototype with a well-regarded second generation online catalog in retrieval from that common set of several million records. We have to exclude records in MELVYL that are not in our local catalog. (We could also compare performance with the different online catalogs on individual campuses to the extent the subset includes records from each campus.)

It became clear that such an evaluation should be gradual, exploratory, and systematic and so we decided to defer systematic evaluation until after the prototype interface had been rebuilt with the additional amenity of a mapping between NGA feature type codes and *Library of Congress Subject Headings*. We will perform and report an evaluation during 2005.

2.3. Ranking evaluation: Some retrieval systems yield retrieval sets ranked with respect to relevance, but online library catalogs ordinarily do not, using simple binary Boolean operations without ranking. We have not attempted to build in ranking, except in one regard that is of special importance in geographic search.

Places can be found by using a single value of latitude and a single value for longitude, commonly a centroid. But dealing with *areas* typically involves overlap between two or more areas and a fuller spatial description is required. Two areas may be mutually overlapping, both occupying the same, identical space; one area may contain (or be contained by) another area; one area may partially overlap another; or two areas may be non-overlapping, each entirely outside of the other (“disjoint”). How are we to compare and rank geographic areas when they can be of any size, shape, and with any degree of overlap (or none) with any other area? Prior treatments had used relatively simple calculations of overlap. We experimented with three enhancements:

1. Convex hulls. Convex hulls are irregular convex polygons, like a tight rubber band around an object that ignores concave areas. (See Figure 13). Convex hulls are an efficient way to represent

an area more accurately using a minimum bounding box with only two values of latitude and two values of longitude;

2. On-shore adjustment. Adjusting for the portion of a coastal area that is on-shore; and

3. Regression analysis. Using regression analysis provides a refined use of the portion of the query area that overlaps the target area; the proportion of the target area that overlaps the query area; and the fractions of each that are onshore.

Analysis showed that these enhancements provided a major increase in area ranking performance. Our paper presenting these findings at the 2004 European Conference on Digital Libraries (ECDL) received the “DELOS Best ECDL Paper Award” award (Larson and Frontiera 2004).

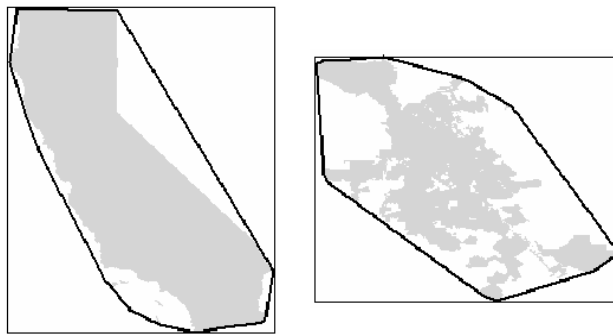


Fig. 13: Minimum bounding boxes and convex hulls for California and San Jose, CA.

Michael Buckland attended the Outcome-based Measurement Workshop, organized by Performance Results, Inc., in December 2002. The two expected outcomes are demonstration of improved techniques, thereby encouraging innovation, and influencing the adoption of best practices, standards and protocols, thereby improving library and museum services' outcomes. Outcomes-based measurement is expected to become progressively more meaningful in subsequent work as these exploratory techniques move closer to operational deployment by libraries and museums.

EXTENDING SEARCHES

The description of the Going Places interface discusses only searching in the local catalog and a local copy of the NGA gazetteer. Extending searches to other library resources will be done, as it was in the ECAI Cultural Atlas of Iraq, by adding a Z39:50 software with a menu of resources supporting Z39:50 service. This will be operational in the revised interface.

ADMINISTRATION

Several people worked on this project. Michael Buckland, Fredric C. Gey, and Ray R. Larson served Principal Investigators. Kim Carl, Aitao Chen, Ruth Mostern, Vivien Petras, and Jeanette Zerneke were active through all or most of the project. David Blundell, Lawrence Crissman, Sarah Ellinger, Matt Meiske, and Patricia Frontiera also contributed. Accounting and administrative support was provided by Jeri Foushee, Charlene Nicholas, and others in

International and Areas Studies, by Kevin Heard and others in the School of Information Management and Systems, and by staff in the Sponsored Projects Office. In addition to IMLS support, the project was also supported by the University of California, Berkeley, DARPA, and other sources.

A project website was established at <http://ecai.org/imls2002/> and several presentations about the project and what has been learned were made at conferences, seminars, and workshops, including IMLS WebWise 2004, in Chicago, the joint IMLS and NSF PIs meeting in Washington, DC.

Geographic techniques need to be evolved and tested in an international context. Project work has been reported at ECAI's twice-yearly international meetings in Vienna, Austria, in April 2003, jointly with the Computer Applications in Archaeology Conference; in Osaka, Japan, in September 2003, jointly with the Pacific Neighborhood Consortium, Electronic Buddhist Text Initiative, the Special Interest Group for Computers and the Humanities of the Information Processing Society of Japan, and the "Minpaku" National Museum of Ethnology; at the conference is Cultural Heritage and Collaboration in the Digital Age with the Pacific Neighborhood Consortium and other groups at Maha Chakri Sirindhorn Anthropology Centre in Bangkok, Thailand, in November 2003; and at the Digital Libraries and Digital Collections in the Global Community conference, jointly with the Pacific Rim Digital Libraries Alliance and the Pacific Neighborhood Consortium at Academia Sinica, Taipei, Taiwan, in November 2004. In May 2004 ECAI organized the first annual congress on cultural atlases in Berkeley and project work was presented as a paper and as a poster. In conjunction with these conferences, workshops on the design and use of gazetteers were conducted in Taiwan in 2003 and at Berkeley in 2004.

Future presentations on the project are scheduled at the Association of American Geographers conference in Denver April 2005 and, we expect, at ECAI's Second International Congress on Cultural Atlases in Shanghai in May 2005, hosted by the China Historical GIS Center of Fudan University. (See <http://ecai.org/Activities/conferences.asp>).

THE BIGGER PICTURE

1. *Gazetteers.* Working on this project has strengthened our conviction that *georeferenced name authority files* -- gazetteers -- are critically important for improved geographical search support. They have three significant advantages over conventional name authority files used in library vocabulary control: Georeferencing provides greater stability than reliance on administrative jurisdiction boundaries; georeferencing facilitates disambiguation; allows enhanced use of spatial relationships; and, especially, georeferencing supports the use of map visualizations, which are an increasingly common feature of personal and professional computing environments.

The importance of gazetteers for library purpose increases the need for more and better gazetteers and for network accessible gazetteer services. Better design includes support for multiple languages, multiple scripts, a wider range of feature types, and improved interoperability among gazetteers and between gazetteers and other resources.

2. *Derived Projects* This project gave us increased confidence and credibility for applying its lessons.

Language maps, a little-known genre, constitute a significant resource for the geography of culture and are an important resource for the study of dialects, diasporas, and endangered

languages. David Blundell and Michael Buckland have extended the Cebuano exercise by continuing to work with collaborators in Taiwan and Australia on language maps for Austronesian languages and dialects, supported by grants from the U.C. Berkeley Shung Ye Museum of Formosan Aborigines Endowment Fund. If funded, a larger project would use link language maps to bibliographical and other resources to provide context for each language.

The Hindi news reports project has been extended in a geo-temporal interface to Russian news reports in Cyrillic script. A problem here is that library transliteration (romanization) standards do not necessarily follow international practice. This work was presented at the ACM SIGIR-2004 Workshop on Geographic Information Retrieval and at the poster session of the European Conference on Digital Libraries, 2004. (See Gey and Carl 2004 and http://ecai.org/samples/Russian/proto-0704/www_RussianNews_by_city.html)

A major derived project is the *Religious Atlas of China and the Himalayas*, under the direction of Lewis Lancaster, supported by the Luce Foundation. Variations in time and place are vitally significant to the practice of religion. During the 5,000 years of Chinese history, Daoism and indigenous practices flourished in every region of the realm. Foreign religions including Buddhism, Zoroastrianism, Islam, Judaism, and Christianity, entered and spread throughout China. These religions are all associated with many locations: temples, pilgrimage routes, sacred spring and mountains, schools, and many more. The Atlas aims to create interactive maps, timelines, gazetteers, and links to images, websites, texts, and datasets. Collaborators in this international project are specialists in Asian art, history and religion; and in geographic information systems and digital library development. The first phase focuses on the compilation of gazetteers of Chinese and Tibetan Buddhist place names, a major intellectual, linguistic, and technical challenge. See <http://ecai.org/chinareligion/>

2. *Place and Time* Place and time are related, since places, being cultural, change with time. Places have temporal aspects and time periods have geographic aspects. This project indicated the importance of supporting search by time as well as by place. People tend to denote time by using named time periods such as *neolithic*, *antebellum*, *Clinton administration*, and so on. Important insights were that time period names can be treated much like place names and that a gazetteer makes a good basis for the design of a directory of named time periods. (See Feinberg 2003).

3. *An new agenda.* Topical search is quite well understood. This project gave us insight not only into improved geographical search but also into how named time periods could be better handled. Meanwhile, biographical dictionaries draw heavily on times and places. These components – What, Where, When, and Who – can move us forwards in providing a structured approach for learners. We are fortunate and grateful that we have received support from IMLS to extend our work in this way in a new project: “Support for the Learner: What, Where, When, and Who”. See <http://ecai.org/imls2004/>

4. *Metadata as Infrastructure.* Working with vocabularies, formats, and standards and their relationships strongly reinforced our belief that metadata systems are as much infrastructure as are networks, hardware, and software. When invited to present testimony to the ACLS Commission on Cyberinfrastructure for the Humanities and Social Sciences we drew on this

project to argue that case. Seeing metadata as infrastructure paves the way for a new era of progress through standardization, interoperability, and resource sharing.

Related publications

- Michael Buckland. *Going Places in the Catalog: Enhancing Scholarly and Educational Resources with Geospatial Information*. In *WebWise 2004: Sharing Digital Resources*, March 3-5, Chicago. Handout: <http://ecai.org/imls2002/WebWise.pdf> Powerpoint presentation: <http://ecai.org/imls2002/webwi04.ppt>
- Michael Buckland. *Metadata as Infrastructure; Interoperability; and the Larger Context*. Testimony to the Commission on Cyberinfrastructure for the Humanities and Social Sciences. Berkeley, Aug 21, 2004. <http://www.sims.berkeley.edu/~buckland/cyber04.pdf> and <http://www.sims.berkeley.edu/~buckland/Cyber8-21-04.PPT>
- Michael Buckland and Lewis Lancaster. *Combining time, place, and topic: The Electronic Cultural Atlas Initiative*. *D-Lib Magazine* 10(5), May 2004. <http://www.dlib.org/dlib/may04/buckland/05buckland.html>
- Aitao Chen. *Notes on Library of Congress Subject Headings (LCSH) Chronological Subdivisions*. Supplementary Report for Institute of Museum and Library Services National Leadership award number LG-02-02-0035-02. Berkeley, March 2004 <http://ecai.org/imls2002/timesubdiv.html>
- Melanie Feinberg. *Application of Geographical Gazetteer Standards to Named Time Periods. Draft Report for Institute of Museum and Library Services National Leadership award number LG-02-02-0035-02*. Berkeley, October 22, 2003 http://ecai.org/imls2002/time_period_directories.pdf
- Fredric C. Gey and Kim Carl. *Geotemporal Access to Multilingual Documents*. Presentation at the ACM SIGIR-2004 Workshop on Geographic Information Retrieval and at the poster session of the European Conference on Digital Libraries, 2004. http://ucdata.berkeley.edu/personal/fred/my_papers/geotemporal_access_ecdl_2004_poster.pdf
- Ray R. Larson and Patricia Frontiera. *Spatial Ranking Methods for Geographical Information Retrieval (GIR) in Digital Libraries*. Paper presented at the European Collaborative Digital Library Conference, 2004. http://cheshire.lib.berkeley.edu/ECDL2004_preprint.pdf
- Religious Atlas of China and the Himalayas*. <http://ecai.org/chinareligion/>
- Vivien Petras. *Statistical Analysis of Geographic and Language Clues in the MARC Record*. Technical Report. Dec 8, 2004. <http://metadata.sims.berkeley.edu/papers/Marcplaces.pdf>