

Submitted for publication in *College and Research Libraries*.

<http://www.ala.org/ala/acrl/acrlpubs/crljournal/collegeresearch.htm>

Accepted for publication in September 2007. Published version may include revisions.

Geographic Search: Catalogs, Gazetteers, and Maps

Rev. July 19, 2006, Feb 20, 2007..

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

Ruth Mostern, Assistant Professor, School of Social Sciences, Humanities and Arts, University of California, Merced, CA 95344

Vivien Petras,

School of Information, University of California, Berkeley, CA 94720-4600

Correspondence to: buckland@sims.berkeley.edu

Acknowledgement: This work was partially supported by Institute of Museum and Library Services National Leadership Grant LG-02-02-0035-02, Oct 2002 - Sept 2004 for “Going Places in the Catalog: Improved Geographic Access,” by DARPA award N66001-00-1-8911, 2000-04, for “Translingual Information Management Using Domain Ontologies,” and by the U.C. Berkeley Shung Ye Museum of Formosan Aborigines Endowment Fund. We are grateful for the advice and assistance of David S. Blundell, Kimberly Carl, Lawrence W. Crissman, Sarah Ellinger, Melanie Feinberg, Patricia Frontiera, Matt Meiske, Ruth Mostern, and Jeanette L. Zerneke.

ABSTRACT

Libraries need to support geographic search. The traditional reliance on place names and political jurisdictions needs to be complemented by greater attention to space, using latitude and longitude. If place name authority files are linked to (or developed into) place name gazetteers, spatial coordinates can be added, places can be located in space, similar and multiple place names can be disambiguated, additional spatial relationships can be established (e.g. near, between). Map visualizations used to display geographic aspects of retrieved sets and also provide a more flexible way in to specify the geographic facet in search queries. Analyses show

that library catalog records contain geographic data that remains unused. Recommendations and prototype interfaces are presented.

INTRODUCTION

Libraries have a broad need to support geographic search: directly for a place (e.g. FIND **Andorra**) or for an aspect of a place (e.g. FIND **Andorra -- Antiquities**), or indirectly through geographic aspects of any topic (e.g. FIND **Folklore – Andorra**). A wide variety of documents may need geographic search support. Searches for some genres, notably socio-economic numeric data series, ordinarily require that a geographic area be specified.

Place, Space, and Area

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. Cities expand, absorbing neighboring places, and countries change both names and boundaries. Space, in the sense of an area specified by physical measurements is a more modern concept and is more stable.¹ Places can usually be characterized by the spaces they occupy 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 geographic searching in library catalogs is provided primarily by using place names as geographic name subject headings (MARC 651) or as geographic aspects of other kinds of subject headings through the use of “geographic subdivisions” (65X|z) qualifying non-geographic subject headings (e.g. 650|a**Folklore**|z**Andorra**). In addition, one can use place names in titles, in notes, or as corporate authors. These approaches are consistent with ordinary discourse, but are 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 do searches relating to that area during that time?)

Library catalogs and the data in them reflect descriptive cataloging rules of great complexity. Most of the geographic indicators in the records are not searchable in current catalogs, and many of the optional forms of description are rarely encountered and so, although valid, might not be used in searches because unexpected (e.g. the Geographic Area Code *c* for *Cold Regions*). Further, the rules are continuously evolving, so large research libraries’ catalogs contain legacy records, some created in the nineteenth century, which may not been revised as the rules change. The growth of geographic information systems and in the use georeferenced data opens new opportunities.²

SAMPLE CATALOG RECORDS

The following two abbreviated catalog records, encoded in MARC, illustrate existing practice. Geographic clues are indicated in **bold** type and we have added explanatory notes in *italic*.

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 of Publication Code (used in field 008), instead of the quite different Geographic Area Code prescribed for 043. It should be e-uk-ui for “Europe. Great Britain Miscellaneous Island Dependencies,” which is legally correct but does not provide the location.*
- 050 00 |a TF764.M27 |b P4 1970 *Geographic code embedded in Library of Congress Classification number*
- 082 00 |a 388.4/6/094289 *Geographic code embedded in Dewey Decimal Classification number. Error: Should be 94279.*
- 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.*
- 610 20 |a **Manx** Electric Railway Company. *Adjective for Isle of Man used in corporate name used as subject heading.*
- 650 _0 |a Street-railroads |z **Man, Isle of**. *Geographic subdivision using inverted form of name.*
- The island known as Man is represented six different ways, two of them incorrect.*

New Amsterdam and its People

- 008 710331r19691902nyuace 000 0 eng *State of publication code for England*
- 043 __ |a **n-us-ny** *Geographic Area Code for the state (rather than the city) of New York.*
- 050 00 |a **F128.4** |b .I58 1969 *Library of Congress Classification number for New York city.*
- 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 *Nickname for New York state 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 two 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 (5XX), including 545 Biographical or historical data.

Library catalog records contain more geographic 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. Strategies for improvement include extending search capability to make better use of the geographic clues already in catalog records, enrichment of record content, and using gazetteers and map displays to augment the use of place names.

EXTENDING SEARCH CAPABILITY

An obvious way to increase geographic search capability would be to support search on the Geographic Area Code (MARC 043). The most commonly found codes are in the form [Continent] – [Country] – [Sometimes, Province or state], so, for example, *e-lu* for Luxemburg and *n-us-id* for Idaho. These are intuitive and easy to use. However, we know of no library catalog that allows search on this code, so adding that capability is an obvious, feasible enhancement. The Geographic Area Code system is, however, vulnerable to criticism. First there is no provision for geographic areas smaller than county or, for a very few countries, province or state. Second, once one moves beyond the continents and the countries in them, the system is complex and less intuitive. The Indian Ocean has hyphenated country divisions, like the continents, but the Atlantic and Pacific oceans do not. There is provision for regions (e.g. *aw Middle East*), border regions (e.g. *m Intercontinental areas (Eastern Hemisphere)*), classes of countries (e.g. *d Developing countries*), geographic 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*). There is, presumably, collection warrant for these codes, but the great majority appear to be rarely assigned and searchers might well not think of using them. A map display of the codes used in each catalog might help searchers understand the available options. In future revision, consideration could be given to using the same geographic coding system in both the 008 and the 043 fields, or, much better, adoption of an international standard code as has been done with the code for languages.

Geographic aspects are commonly incorporated within classification numbers, as in the sample records above. Using these codes would depend on the geographic element being an identifiable facet, which would be straightforward with the Universal Decimal Classification (UDC), but that system is very rarely used. With the Dewey Decimal Classification and Library of Congress Classifications, the two most widely used systems, the geographic component would need to be identified and marked to be usable. A map display, as a prompt, would help.³

CATALOG RECORD ENRICHMENT

We had expected that many records would have geographic name subject headings or subdivisions (651 or 65X \$z) but not geographic area codes (043), or vice versa. It should be feasible to enrich records algorithmically by generating geographic area codes from the geographic name subject headings or subdivisions. Likewise, 043 codes could be used to generate geographic subdivisions to enrich the Subject Headings, which would be more useful since Subject Headings are already searchable. However, we found little scope for such enrichment in the records we examined.

A set of 5,065,574 MARC records from the union catalog of the University of California libraries, kindly made available for our research by the California Digital Library, was analyzed. All of these records originated from the Library of Congress, but had call numbers and possibly other modifications made by University of California campus libraries. We found that records with geographic 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. So there is little scope for catalog record enrichment by inferring missing 043 Geographic Area codes from Geographic

Name Subject Headings (651) or Geographic Subdivisions (65X|z) or vice versa. It is a welcome conclusion. The situation is better than we had expected.⁴

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 places in which the language is used. This is, we assume, generally true in the context of U.S. academic library collection development policies. The inference is that in collection development, books in English are preferred and then, to strengthen holdings about specific foreign countries, titles published in that country are added selectively and these foreign books tend to be in the language of that country. So within library collections foreign languages and foreign places of publication can serve as imperfect but serviceable indicators of geographic scope. These conclusions may be true of publishing practices generally, but our evidence relates specifically to library collections, which is what catalogs represent.⁵

GAZETTEERS FOR PLACE NAME CONTROL

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 ought to 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, and sometimes the names are used metaphorically, not geographically. 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, less reliably, by natural language parsing techniques. (One simple rule is that a word not in a standard dictionary is probably a proper name for a place, a person or an institution).

A second major problem is lack of connection between library place name authority practice and place name gazetteers, the well-established authority genre developed for 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 gazetteers ordinarily include at least the following elements:

- The place name;
- The country or area within which the place is located;
- Geographic description codes, commonly known as "feature types": a coding for the kind of place, e.g. castle, lake, inhabited place, airport, etc.
- Spatial references: latitude and longitude, usually a single point defined by the intersection of a single line of latitude and a single line of longitude, but potentially more complex; and
- References to or from other names for the same place.

When used at the back of an atlas the gazetteer also serves 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 named place can be found. But the gazetteer is an important 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, previously the National Imagery and Mapping Agency (NIMA)) and the Board on Geographic Names (BGN) at <http://www.nga.mil> and 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 at the University of California, Santa Barbara (<http://middleware.alexandria.ucsb.edu/client/gaz/adl/index.jsp>).

Geographic feature type codes. Gazetteers also hold the promise of supporting more precise searching for specific kinds of places because each entry contains a “feature type” code indicating what kind of place it is. A library catalog could take advantage of these assets if it were to use a gazetteer instead of (or to augment) standard name authority files. This idea becomes the more feasible with the emergence of protocols for interrogating online gazetteers, such as the Alexandria Digital Library Gazetteer Server Protocol.

Gazetteers have been developed primarily for contemporary geography and for governmental, industrial, and military needs. For the humanities, social sciences, and libraries an effective 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 each place name was in use, and support for multiple languages and multiple formats.⁶

Comparison of the NGA Geographical Description Codes (GDC) with Library of Congress Subject Headings (LCSH) reveals differences in style and in emphasis, as well as scope and scale, with some 600 NGA GDC codes to over 150,000 LCSH. Sometimes LCSH has greater detail, especially for kinds of historic sites; sometimes NGA has more detail, for example in submarine geomorphology. NGA is only concerned with physical features, so, for example, “School” means a school building, not an institution. LCSH uses the plural form for objects, so NGA “School” corresponds to LCSH “School buildings.” Modest changes to the NGA Geographic Description Codes could both increase interoperability with LCSH and make them more useful in for the humanities and social sciences. After harmonizing singular and plural forms and few variant spellings, there is a direct match or an acceptable synonym in most cases. Mapping the GDC with LCSH provides a basis for linked searches in the form: Find literature about lighthouses [in the catalog] and identify the location of local lighthouses [in the gazetteer].

The special advantage of a gazetteer is the inclusion of spatial coordinates. Latitude and longitude provide a stable framework within which named places can be positioned and found using map visualization software. Hitherto, georeferencing has usually been in point form, the intersection of one line of latitude and one line of longitude, but as the use of map interfaces increases we can expect the use of more complex georeferencing: using two lines of latitude and of longitude to form a bounding box and more elaborate polygons approximating actual boundaries.

Our conclusion is that 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.

MAPS AS CATALOG INTERFACES

Effective geographic searching needs a map display, and a map display interface to a library catalog needs to be time-sensitive for three reasons:

1. Date of publication is a significant variable in searching and sorting. So one should be able to set a map display to any searcher-defined period.
2. Historical search topics are commonly for books concerning a topic at some past period of time. The ability to zoom in time as well as space would be useful for catalog records encoded for a past period (with a chronological subdivision) in addition to geographic scope.
3. Maps allow successive boundaries and place names to be overlaid on to a stable geographic framework, so that we can see through the transient political changes to the more stable cultural and geographic features.

The time-based interactive mapping (Time Map TM, www.timemap.net) was used in three technical exercises:⁷

A Map Interface to Catalog Records.

A map-based search interface was created to some 700 books about, or published in, the Cebuano region of the Philippines. First, catalog records were harvested from online library 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/>. These records were formatted into a tab-delimited spreadsheet. A Cheshire script added georeferencing by searching a Cheshire database of the NGA gazetteer names for the Philippines (over 70,000 entries) for each place of publication and geographic name subject heading or geographic subdivision in the records.

A dynamic map interface supports panning, zooming, and adjusting the time period of interest. Small squares on the map marked places contained in subject headings and small circles indicated places of publication. Clicking on any square or circle would display the corresponding catalog records associated with that place.

The interface also shows contextual information, with map layers for 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). Most of the component data sources were freely available on the Internet, with the major exception of the scanned language maps (kindly provided by David Blundell, then at the National Taiwan University, and Lawrence Crissman, Griffith University, Australia). This interactive interface is at www.ecai.org/imls2002/cebuano/CebuanoIndex.html.

A Map Interface to Documents

Our next interface provided access to actual texts, news reports, which are very much about time and place and so made a good test genre for our purposes.⁸ Place names were extracted from Hindi news documents and the location (latitude and longitude) for each place name was retrieved from a gazetteer of the region and depicted as points on a map. A timebar below the map can be adjusted to limit or expand the period to be displayed and the map shows data only for documents whose dates fall within that range. The time-bar can be used to limit the temporal range of the search. Clicking on a single point or by tracing a bounding box around a number of points defines the place or area of interest. Very brief bibliographic records for the

documents that fall within the spatial and temporal limits of the search are displayed in an attribute table. Selecting one of the records then clicking on the link button will display the full text. See <http://ecai.org/imls2002/hindi/hindimapspace.html>.

A Map Interface and Live Library Searches.

The *ECAI Cultural Atlas of Iraq* is a portal to Iraqi antiquities.⁹ One component is a TimeMap interface allowing one to pan and zoom into any desired area and adjust the sliding time bar to specify a time period of interest. Only data pertaining the area shown *and* the specified date range will be visible on the map display. Clicking on any point of interest will display a spreadsheet providing the name of the site, latitude, longitude, feature type, and, among other data, the link to a separate webpage about that site. Each site page has the customary kinds of links to resources, but the first section, labeled “Library Books,” contains three links which generate live searches using the CHESHIRE search and retrieval system. The three searchable resources are the MELVYL union catalog of the hundred libraries of the ten campus University of California; COPAC, the online catalog of the all the largest research university libraries in the United Kingdom; and the online catalog of the Library of Congress. 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.”

A link that generates a live search ensures that the retrieved set is as up-to-date as the catalog itself and is thus importantly different from accessing a compiled but obsolescing bibliography stored at a website. Increasingly, a retrieved catalog record contains a link to the text or image cataloged. In this way a website can provide access to libraries’ documents as well as to catalog records.

THE “GOING PLACES” INTERFACE

To explore geographic search support further we developed a prototype interface able to:

- Search an online catalog;
- Search an online gazetteer;
- Pass catalog data to the gazetteer and vice versa;
- Generate map displays of the geographic aspects of retrieved sets; and
- Use map displays to help form the geographic aspect of search queries.

Search starting from the library catalog

A search can be initiated by typing a one-word query into the query box for a catalog search, which will search the local catalog (or other sources, if selected from the pull-down menu). By using the several million catalog records kindly made available by the California Digital Library as our “local catalog” we were able to make use of any and all parts of the MARC records. The catalog documents found through the search are displayed very briefly by title in the lower left display. Fig. 1 shows the result of search using the keyword **Folklore**. The tabs above the Display result box can be clicked to generate other arrangements of the retrieved set: by year (date of publication); by Library of Congress Subject Heading; by geographic “Scope” (using 043 geographic area codes); or by LC Classification number.

Clicking on any record in the Display result box will open up a new window displaying the full MARC record of the selected document.

The interface analyzes the Geographic Area Codes (043 field) in the retrieved records and uses a copy of the National Geo-intelligence Agency gazetteer to show, in the map box, which countries are represented in one or more books. 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.

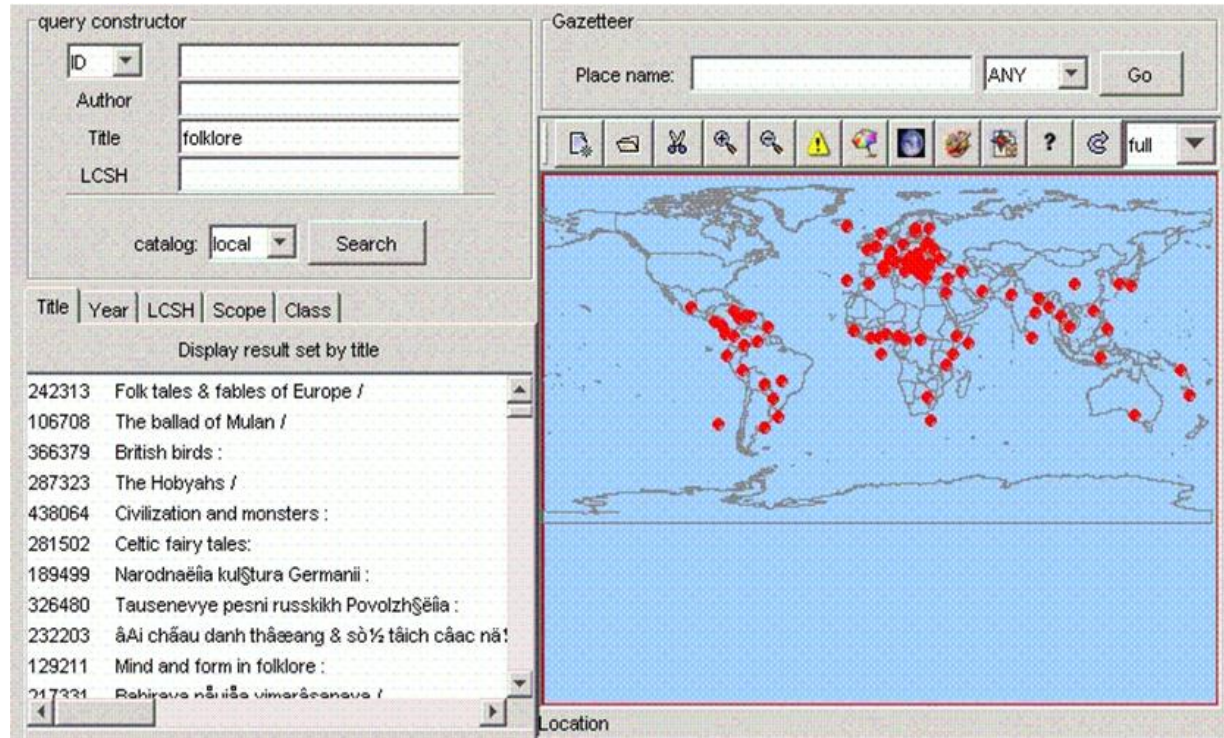


Figure 1. Map display of geographical area codes for a retrieved set of books on Folklore.

Geographic search from within 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 listing the place names found and changing the map display to show dots indicating all the places with that name. 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 Gazetteer.

A search can also be initiated from the map interface itself or from the gazetteer search box. Fig. 2 illustrates a search for capital cities in South America performed by selecting the feature type "capital city" ("PPLC") in the pull-down menu for NGA Feature Designation Codes and using the cursor to drawing a bounding box around South America. The results are displayed both in spreadsheet window and also by red dots in the map interface.

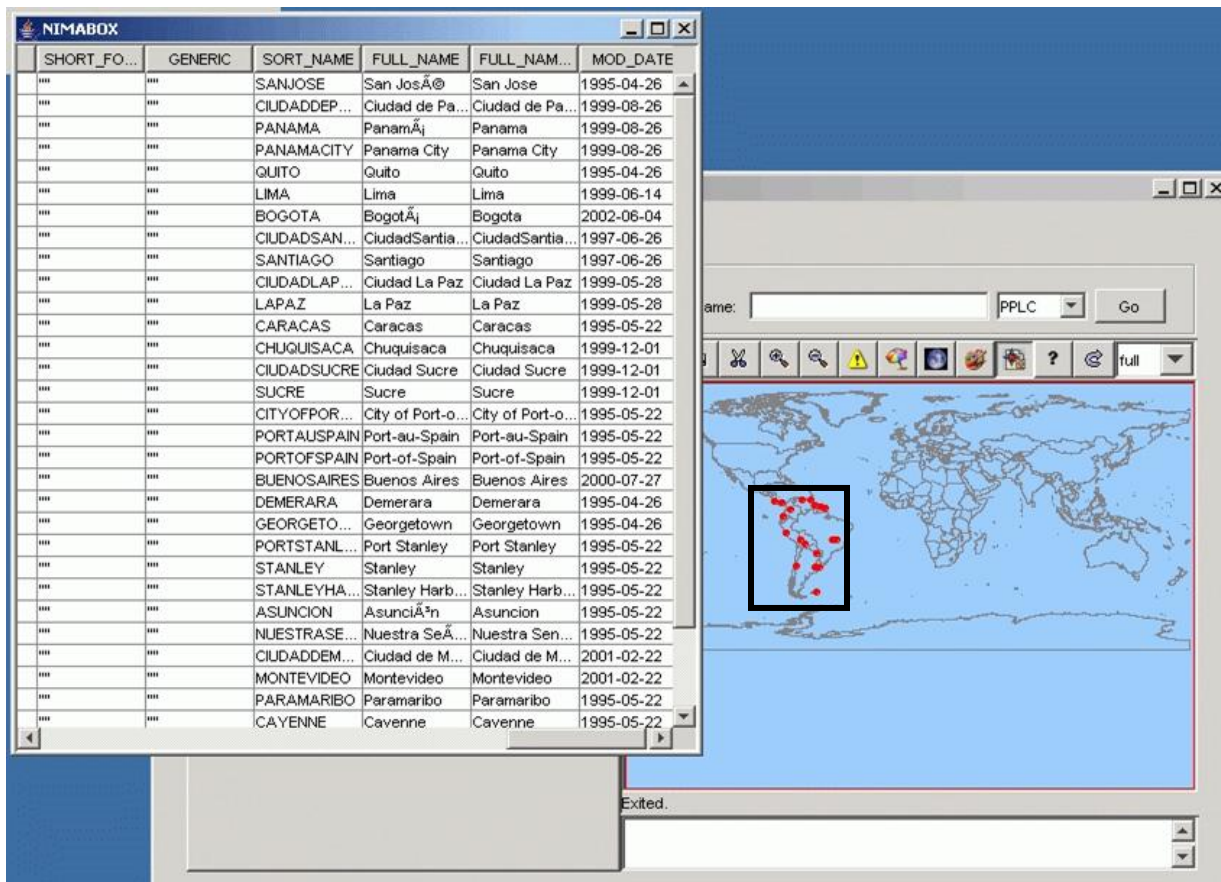


Figure 2. A bounding box drawn around South America and the Geographic Description Code for capital cities formed a query to a gazetteer to retrieve a list of South American capital city names for a search in a library catalog.

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 designed, searches in the gazetteer generate place names as queries to be searched in the catalog. Explicit use of the NGA feature designation codes (such as PPLC) for catalog searching would be possible with a mapping between NGA Feature types to corresponding LC Subject Headings.

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 user interface itself. The interface uses Java and so can be used from any web browser without downloading software. The experimental prototype was intended to demonstrate proof of concept. The software is complex and only suitable when high bandwidth is available. Development and maintenance has been discontinued and a simpler replacement design is in preparation.

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 part of the North American continent. But searchers can also be expected to be interested in other spatial relationships, notably *near* (e.g. within a hundred miles of a place), *between* two places, and *overlapping* another area. These kinds of spatial relationships can be calculated if the latitudes and longitudes are known.

We implemented *between* and *near* 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. The smallest rectangular bounding box that includes two points gives a simple definition of the area and, therefore, places *between* the two points. For *near* a square bounding box centered on a point gives a simple, but economical interpretation of near. How near depends on the size of the box. Defining *near* more precisely becomes computationally very intensive. A bounding circle centered on a location would provide an exact expression of “within x miles” of a point. Trigonometry could be used to determine all the combinations of latitude and longitude that are within a circle of any given radius of the point of interest. The gazetteer would then have to be combed for all the places located within that radius. Both operations require excessive computation and are not scalable. A far less demanding approximation can be achieved by dividing the earth’s surface into many small rectangular tiles, each one with an associated subset of the gazetteer containing the place names within that tile’s area. A bounding circle can then be approximated by determining which *tiles* are within the circle, then taking all of the places within all of those tiles.

SEARCH PERFORMANCE

It appears that any individual place that 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 geographic 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 geographic qualifier, and a broader entry for the feature type, with a geographic subdivision. So, for example, a particular castle in Austria should be assigned two headings, e.g.:

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

The combination of a specific instance under its name and the category (castles in Austria) is important. If you knew of Halbturn castle you could find it if you knew to search under Schloss Halbturn, and a subject keyword search on “Halbturn” ought include it in its retrieved set. Halbturn castle should also be findable by tediously scrolling through the result of the broader search on “Castles – Austria”. The geographic subdivision is not precise enough to distinguish castles in an area smaller than Austria, such as Burgenland province where Halbturn is located. This approach is especially inefficient for castles in borderlands, especially where borders and country names have been unstable. So, for example, 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 another model heading in Library of Congress *Subject Cataloging Manual: Subject Headings*, section H 1334:

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

(Actual records in the Library of Congress catalog, however, do provide narrower geographic subdivisions than the manual prescribes.)

The Going Places interface could support geographic searches more precisely. Drawing a bounding box (or circle) on a map interface allows one to define geographic searches in terms of small areas of interest. 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.

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 provide ranking, except in one regard that is of special importance in geographic search.

Places can be located by using a single value of latitude and a single value for longitude, commonly a centroid. But *areas* typically involve overlap between two or more places 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 are irregular convex bounding polygons that ignores concave areas, like a tight rubber band around an object. Hulls are a more exact way to represent an irregular area than a rectangular bounding box based on only two values of latitude and two values of longitude.
2. Adjusting for the portion of a coastal area that is on-shore; and
3. 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.¹⁰

THE BIGGER PICTURE

1. Georeferenced name authority files -- gazetteers -- are critically important for improved geographic 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.

Georeferencing is now being added to place name authority records, which is a significant improvement. Nevertheless, using or linking to a fully formed gazetteer record is likely to provide richer detail and to facilitate the updating of names, feature types and boundaries.

2. 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, catalogs, and other resources.
3. Places, being cultural, change with time, increasing the importance of supporting search by time. In practice people discuss time in terms of named time periods: *neolithic*, *antebellum*, *Clinton administration*, and so on. Named time periods can be treated like named places and gazetteers make a good basis for the design of a directory of named time periods. Time and place are related, since places have temporal aspects and named time periods have geographic aspects. Chronological subdivisions can be used, of course, but much more could be done.¹¹
4. Vocabularies, formats, standards, and their relationships infrastructure as much as networks, hardware, and software are.
5. The description of the Going Places interface discusses only searching in a local catalog and a locally held gazetteer. Searches can be extended to other network-accessible resources by adding a Z39:50 software with a menu of resources supporting Z39:50 service.

Place, like time, is a very basic and important concept. Libraries have an opportunity to transform support for geographic searching by connecting resources that have long existed in isolation, but now become interoperable in a digital library environment.

References

¹ Michael R. Curry, "Toward a geography of a world without maps: Lessons from Ptolemy and postal codes," *Annals of the Association of American Geographers* 95, no 3 (Sept 2005): 680-91.

² **Error! Main Document Only.** Special Issue on Georeferencing and Geospatial Data, *D-Lib Magazine* 10, no. 5 (May 2004). Accessed July 19, 2006.
<http://www.dlib.org/dlib/may04/05contents.html>.

³ A map prompt for physical zone codes in the Universal Decimal Classification can be found in Helmut Arnhold, Helmut. 1963. *Geographie und Nachbargebiete: DK-Auszug*. Bücherei des Dokumentalisten, 23. (Berlin: Deutsche Akademie der Wissenschaften zu Berlin, Institut für Dokumentation, 1963.

⁴ Vivien Petras, *Statistical Analysis of Geographic and Language Clues in the MARC Record*. Technical Report, 2004. Accessed July 19, 2006.
<http://metadata.sims.berkeley.edu/papers/Marcplaces.pdf>

⁵ Petras, *Statistical Analysis*.

⁶ *ECAI Gazetteer Project*, 2002. Accessed July 19, 2006.
<http://ecai.org/projects/gazetteer/index.html>

⁷ Michael Buckland, Fredric C. Gey, & Ray R. Larson, *Going Places in the Catalog: Improved geographic Search. Final Report*. (Berkeley: University of California, School of Information Management and Systems, 2004). Accessed July 19, 2006. http://ecai.org/imls2002/imls2002-final_report.pdf

⁸ Gey, Fredric C. 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. Accessed July 19, 2006. <http://ucdata.berkeley.edu:7101/staff/gey/papers/geotemporal-access-ecdl-2004-poster.pdf>

⁹ Electronic Cultural Atlas Initiative, *ECAI Iraq*. Accessed July 19, 2006. <http://ecai.org/iraq>.

¹⁰ 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. Accessed July 19, 2006. http://cheshire.lib.berkeley.edu/ECDL2004_preprint.pdf

¹¹ Vivien Petras, Ray Larson, & Michael Buckland, Time Period Directories: A Metadata Infrastructure for Placing Events in Temporal and Geographic Context. *Forthcoming in: Opening Information Horizons: Joint Conference on Digital Libraries (JCDL)*, Chapel Hill, NC, June 11-15, 2006. Accessed July 18, 2006. <http://metadata.sims.berkeley.edu/tpdJCDL06.pdf>. Also Support for the Learner: Time Periods. Accessed July 19, 2006. <http://ecai.org/imls2004/timeperiods.html>