

GOING PLACES IN THE CATALOG: IMPROVED GEOGRAPHICAL ACCESS

[EXCERPTS FROM DRAFT GRANT PROPOSAL]

Submitted by

Michael K. Buckland, Fredric C. Gey, and Ray R. Larson
Electronic Cultural Atlas Initiative and School of Information Management & Systems,
University of California, Berkeley

SUMMARY

Library catalogs are well-designed to support searching by author, by title, and by topic. But support for finding material relating to places is currently quite limited. One can search for place-names in titles or in subject headings, or specify suitable place-related classification numbers, if known. In addition to the basic problems of using words and classifications numbers in searches, there are some additional difficulties when the interest is in a geographical location rather than in a geopolitical unit, such as a country. Fortunately, places have a unique advantage: Unlike authors' names, titles, and topics, places can be defined unambiguously by their coordinates of latitude and longitude. Further, since gazetteers and geographical (map) displays are becoming available for use in a digital networked environment, the tools for putting latitude and longitude to use are at hand.

A four-point strategy is proposed to take advantage of new and existing (but neglected) resources to enhance geographic searching:

- (1) Existing library records: Better use can be made of data already in library catalog records;
- (2) Gazetteers: Linking online catalogs with online gazetteers could transform geographical searching;
- (3) Maps: Search results can be visualized and new kinds of query supported with map-displays; and
- (4) Resources elsewhere: Searches in the libraries' collections can be extended to find other resources relating to the same locality.

This research project will show how place-related searching in library catalogs can be substantially improved and how place-related searches can be extended to scholarly and education resources in and beyond library catalogs.

NARRATIVE

INTRODUCTION AND NATIONAL IMPACT 2

Geographical searching now 2

Why geographical searching offers a special opportunity 3

Why geographical searches matter 4

ADAPTABILITY - INNOVATION WITH FAR-REACHING RESULTS 4

Resources 5

DESIGN OF THE RESEARCH PROGRAM 5

INTRODUCTION AND NATIONAL IMPACT

The interests of library users often relate to places. They want to know about hiking in Nepal, the history of Japan, French cuisine, Cape Town's Table Mountain, the trade balance with China, birds in New Zealand, and so on. Effective searching by place is a function that librarians do need to provide.

Geographical searching now

At present, searching for places in library catalogs is supported in the following ways:

1. Classification numbers include geographical codes when appropriate. A travel guide to the San Francisco Bay Area might be classified at F 868.S156.3 in the Library of Congress Classification or at 917.94604 in the Dewey Decimal Classification.
2. Subject searches for place-names as main-heading (MARC 6XX \$a) e.g. 651 \$a Albania.
3. Keyword searches for place-names within subject-headings (MARC 6XX \$z) e.g. 650 \$a Libraries \$z Albania
4. Title keyword searches for place-names in titles, e.g. *A Journey Through Albania*.

There are significant problems in the existing approaches.

Classification numbers. Using classification numbers to search by place is feasible. A powerful example is Beghtol's Experimental Fiction Analysis System (EFAS), which incorporates geographical subdivisions from the Universal Decimal Classification (Beghtol 1994). Nevertheless, searching by place using classification numbers is problematic. The standard geographical divisions in classification schemes provide only for larger and/or well-known places. They tend to use political jurisdictions (countries and major cities) whose boundaries may be unstable over time. This is acceptable for current affairs and political topics, but not for scientific, historical, or strictly local inquiries. The geographical aspect is commonly embedded in arcane ways: It is the 7.946 within the 917.94604 that denotes the San Francisco Bay Area and, in an EFAS example, in a classification code for a character in James's *The Turn of the Screw* it is the 21 in *jst(133jst(133?)jrtjqv(056)jqt(042)jpv(21)jpp(649.1)jo(Miss Jessel)* which associates her with the British Isles. To search by place using classification numbers, you need to know how to specify the place you want. Online catalogs are currently more likely to support search by call number than by classification number, let alone the geographical component with the classification number within the call number.

Subject Headings. Place-names are used as subject headings and also as subheadings, and they also tend to

be geo-political rather than purely geographical entities. (See *AACR2R* chap 23.) Being place-names, they share both the convenience of verbal descriptors and the disadvantages of natural language:

- Ambiguity: *Vienna, VA* or *Vienna, Austria*? *Alameda (city)* or *Alameda (county)*? Galicia (region of Spain) or Galicia (region of Poland)?
- Different transliterations from non-Roman scripts: *Peking* is a variant of *Beijing*.
- Different names in different languages: *Deutschland*, for example, is also known as *Allemagne* and as *Germany*. *Cluj*, in *Romania / Roumania / Rumania*, is also called *Klausenburg* and *Kolozsvár*.
- Names change: *Bombay* is now *Mumbai*. *St Petersburg* became *Leningrad*, but is now *St Petersburg* again.
- Anachronisms: *Germany* did not exist as a country before 1870. *Poland* ceased to be a country in 1795, when divided between *Austria-Hungary*, *Prussia*, and *Russia*, but reappeared in 1918. *Prussia* is no longer a country. *Austria-Hungary* became a much smaller *Austria*. The boundaries of each of these countries have been very unstable over time. Political changes disrupt place-name stability: Witness Napoleonic Europe, the end of the *USSR*, and the Balkans.
- Footprint: Even when a place's name is stable, the area it denotes may not be. Notoriously, cities expand over time, absorbing places around them.

Place-names in Titles, also being verbal, share the problems that geographical subject headings have, but are much less likely to be in a standard form and a relevant place-name is not always provided in a book's title.

Why geographical searching offers a special opportunity

We noted above some difficulties in searching by place. They are more or less similar to the problems in searching for topics, persons, institutions, and events. So why focus on places? Places are uniquely different because, unlike topics, persons, institutions, and events, places have a system for objective specification: latitude and longitude. Further, there is a well-established tool for linking *place-names* with *places*: the gazetteer, most familiar as a list of place-names printed in the back of atlases, serving as an index to the maps.

A gazetteer is, minimally, a mapping of place-names to places, in which places are specified by coordinates: latitude and longitude, e.g. *Washington, DC -- Latitude: 38.90505 North -- Longitude: 77.01616 West*. In this case the "place" of the "place-name" *Washington, DC* is specified by a point where a line of latitude and a line of longitude intersect, commonly the geographical center or the population centroid. A better specification could be provided by giving two values of latitude to define the northernmost and the southernmost tips of the city and two values of longitude to define the western and eastern extent, a "bounding box." An even more precise specification would be a series of values pinpointing with as much precision as desired the boundaries of the city - a "polygon."

Gazetteer entries ordinarily contain other elements in addition to a place-name, latitude, and longitude:

- (i) The "Feature Type," an indication of what kind of place it is: city, lake, church,...;
- (ii) An indication of which larger region or country it is in; and, depending on the intended purpose,
- (iii) Additional data-elements, such as a reference to a map showing that place-name, and, occasionally,
- (iv) The date when a particular name, feature type, spatial coordinate, or relationship was current.

A gazetteer is essentially a set of entries composed of a *place-name* and its *geographical coordinates*. When used as an index at the back of an atlas, each entry also has the address of a map containing that place-name. In a digital gazetteer one could go in either direction: From place-name to coordinates; or from coordinates to place-name. We seek to open up powerful new possibilities by linking gazetteers and catalogs.

1. A gazetteer can be a place-name authority file for vocabulary control purposes: Different place-

names with the same coordinates are deemed to be the same place; different places with the same name have different coordinates.

2. More generally, the coordinates in the gazetteer allow spatial relationships between named places to be calculated: How far is Ann Arbor, MI from Urbana, IL? What cities are within 50 miles of Ann Arbor, MI?

3. Having identified in the gazetteer a place-name associated with a place, that name could then be forwarded to the online catalog as a new search query: What do you have about the following cities located within 50 miles of Ann Arbor, MI?

4. Associating coordinates with place-names allows for map-based search aids: Search results can be presented geographically in a map display. In the other direction, an area on a map could be translated into a list of all the place-names with coordinates within that area.

5. Latitude and longitude are a *lingua franca* across all disciplines and all countries for specifying places. Once the library catalog can translate its place-names into coordinates, its ability to extend a search beyond itself and beyond other library catalogs to other network accessible resources is enormously enhanced.

6. The use of temporal metadata in gazetteer standards makes gazetteer-based searches even more powerful by specifying time ranges as well as place names.

Although we speak here of linking *an* online catalog with *a* gazetteer, we envisage an environment in which online catalogs (and bibliographies) will interoperate with multiple network accessible online gazetteers, and with other resources which become searchable *because* of data made available from a gazetteer.

Why geographical searches matter

The most fundamental advantage of the emerging networked knowledge environment is that it provides the *technological* basis for *sharing* resources of all sorts from all sources. The effect of this is to increase the premium on effective *bibliographical* access. Place, along with time, topic, and creator, is one of the fundamental components in how we define things and search for them. In addition, place is pivotal for interdisciplinary inquiry: Archaeologists, anthropologists, botanists, civil engineers, economists, epidemiologists, geologists, historians, manufacturers, military strategists, sociologists, and others form their specialized communities and domain-specific terminology. But, since they are all more or less interested in space and place, and in spatial changes over time, bringing together everything associated with a particular place at a given time is a particularly effective way to understand the relationships between them and the broader context. This is critical for serious collaboration across different disciplines.

The library catalog, traditionally the “key to the collection,” has been evolving into a broader role as a gateway to any accessible resources, local or otherwise (Norgard et al., 1993). This change increases the importance of the library catalog itself and it makes the search effectiveness of the library catalog even more important. As the range of resources that are searchable increases, the benefits of improved search capability are correspondingly extended.

The objective is to show how, in the emerging network environment, librarians can improve their ability to support geographical searching, thereby significantly enhancing the scale, quality, and utility of the service they provide.

ADAPTABILITY - INNOVATION WITH FAR-REACHING RESULTS

The insight that improved support for searching by place was both needed and feasible arose in part from our work on an IMLS National Library Leadership Project entitled *Seamless Searching of Numeric and Textual Resources*. The final part of that project called for extending searches from text found in textual databases to topically related statistical data in socio-economic numeric databases -- or vice versa. We found that socio-economic numeric datasets nearly always have a geographical aspect that is important to the

meaning of the data and that the obvious approach of using place-names to denote that geographic aspect was unsatisfactory in practice. However, converting the name of a geographical area into spatial coordinates as a bridging device between different place-name systems not only provided a *lingua franca*, but also enabled the use of map displays as a search aid.

Another ingredient in the formulation of this proposal lies in the experience of the Electronic Cultural Atlas Initiative (ECAI), which originated when sixteen scholars met at Berkeley in 1997 to discuss the problems of using digital technology in cultural studies. Religious and cultural developments can only be understood in the context of what else has been happening at the same time and place. Descriptive metadata of intellectual content such as subject headings are discipline or culturally based and subjectively assigned. If only scholars in diverse disciplines could be encouraged to code their data formally with latitude, longitude, and time, they could map and understand their own data better. More importantly, they could also juxtapose their own data with data from other sources and see spatial and temporal relationships that would otherwise go unnoticed. With support from the University of California, Berkeley, a center was established under the Dean of International and Area Studies to initiate an international collaborative effort to transform humanities scholarship through increased awareness of the benefits, in a digital environment, paying more careful attention to place and time. Professor Lewis Lancaster, a pioneer of digital versions of ancient Asian religious texts, is Director. Professor Michael Buckland, with extensive library experience, is Co-Director. With assistance from the Lilly Foundation, the California Digital Library (University of California), and other sources, ECAI is nurturing a rapidly expanding community of some 800 scholars (300+ projects) worldwide in a wide range of disciplines, all seeking to make their data shareable. The Academia Sinica (Taiwan), The British Library, the Arts and Humanities Data Service (UK) are among the well-known collaborators, but, mostly, it is individual scholars and small teams working on their own small projects on a wide range of cultural, social, and historical topics. Numerous specialist committees are fostering standards and collaboration by area and by theme, e.g. trade-routes, cities, religion, sacred sites, etc. An online catalog of these resources, the ECAI Metadata Clearinghouse, is being built. Needed standards and software being developed. (Lancaster & Bodenhamer 2002)

Since software for dynamic map displays from distributed resources - showing change through time - was not available, ECAI has supported development of the TimeMap software being developed at the Archaeological Computing Laboratory, University of Sydney (Johnson 1999). Existing gazetteers, developed for governmental purposes and environmental research are inadequate for the humanities. Important needs include support for place names in multiple languages, non-Roman scripts, greatly extended feature type metadata, and time-coding for every place-name and feature type indicating *when* it was applicable. An NSF Information Technology Research program grant is funding ECAI to develop improved content and format standards for gazetteer entries. ECAI is working on this in close collaboration with the Alexandria Digital Library project in Santa Barbara, Academia Sinica in Taipei, and interested parties elsewhere.

Resources

The research team is exceptionally positioned to undertake the proposed research. It has a file of 10 million MARC records made available to it for precisely this kind of research by the California Digital Library. Prior Federal research grants supported purchase of computing resources, disk storage, and software sufficient to manipulate these records and create indexes on *any* MARC field. We have copies of the Getty *Thesaurus of Geographical Names* and the Census Bureau's *U.S. Gazetteer*. ECAI has established collaborative relationships with the Alexandria Digital Library, which maintains a leading online gazetteer, and with Academia Sinica which is building one.

DESIGN OF THE RESEARCH PROGRAM

The first year has four Objectives:

Objective 1. Demonstrate that additional search capability is feasible if better use were made of existing data. MARC records contain, or may contain, a number of geographical clues which could be used.

MARC field 043 for the geographical scope of subject content, in which the *MARC Codes for Geographical Areas* are used, such as *e-lu* for Luxembourg and *n-us-id* for Idaho.

MARC fields 008/15-17, 044, and 260\$a represent the place of production or publication

MARC fields 008/35-37 and 041 indicate the language(s) of the document.

We have yet to encounter an online catalog that supported searches on 043, 008/15-17, 260\$a, 041, or 044. Yet these fields offer valuable clues. A book on folklore written in Slovene and/or published in Slovenia will tend to be about Slovene folklore, even though the subject headings may not have a geographic subdivision indicating so.

During Year One we will create indexes for these fields to our 10 million MARC records; test the hypothesis that fields 008/15-17, 008/35-37, 060\$a, 041, 043 and 044 are statistically useful for geographical searches and look for other potentially useful correlations. For example, we would expect that *statistically* there is a usable correlation between place of publisher, language, and the geographical scope of a book, but we have no empirical evidence. The first Objective is to test this kind of hypothesis using statistical experiments on our 10 million MARC record file and to design probabilistic search algorithms that exploit these neglected data. (A by-product: These algorithms should also be useful for enriching existing catalogs by assigning lacking data to improve the completeness of the catalog records, e.g. the presence of the geographical subdivision *Luxembourg* in 6XX\$z in a record implies that field 043 should be encoded *e-lu* and vice-versa.)

Objective 2. Demonstrate improved place-name authority control by coupling an online library catalog with an online gazetteer. Gazetteers can provide explicit place-name authority control by mapping variant and successive place names. (This should be implicit from the latitude and longitude when it is not explicit.) Linking a catalog with an online gazetteer adds *geo-referencing*, that is, it relates place-names to latitude and longitude. Geo-referencing is important because it provides place-name vocabulary control and a basis for extending searches to resources beyond libraries.

In Chinese *Beijing* means “Northern Capital” and eighteen different places have been named *Beijing*. They are most clearly differentiated by their different latitudes and longitudes, and, if known, by the time when used.

Pekin, *Pei-ching* and *Pekine* are additional variant spellings of *Beijing*. They can be recognized as being synonyms if they have the same latitude and longitude. *Peiping* is a different word, meaning “Northern Peace,” but it used in Taiwan to refer to *Beijing* without calling it a capital.

Latitude and longitude allow stable reference points. The names of places change, and what a name refers to is not always stable. For example, the boundaries of *Germany* and *Poland* have varied greatly. Specifying latitude and longitude is the only way to specify a location reliably through time. The second objective will be to demonstrate how an online gazetteer can be used to improve place-name authority control.

Objective 3. Demonstrate how geo-referencing would allow an online catalog to display search results in map displays, giving library users a new way to understand search results. GIS technology allows structured data with spatial coordinates to be visualized as “map layers” in a common coordinate system. Many map layers may be superimposed in a single environment for viewing and analysis, so that topography, linguistic zones, and trade routes, for instance, could all serve as elements of a “base map” to contextualize the results of spatially referenced bibliographic searches. Figure 1, shows the ECAI search interface which accepts either place names or coordinates (latitude and longitude) as queries. Figure 2 shows the results of search query “Australia,” with each item (image, text, etc.) represented as a symbol appropriately placed in a map of Australia. The table lower left shows some of the metadata for the same items.

Objective 4. Demonstrate how searches can be extended through the use of feature types. Gazetteer entries

ordinarily include a “Feature Type,” metadata indicating what kind of place it is: City, lake, county, etc. An online gazetteer should be searchable by feature type. For example, if a searcher searches for all “cities” in a given region, she could generate place-names to use in new catalog searches. Feature types could also be used for more complex spatial queries, such as “What books on travel and description were published in the 1950s concerning places within 25 miles of dams in California?”

The second year has the following objectives.

Objective 5. Demonstrate the use of interactive map displays as a geographical search enhancement. Maps can also be used to define or refine searches, e.g. by clicking on a pre-defined region. One could click on one or more pre-defined areas (in “base-maps”) or allow the searcher to draw an arbitrary region to mean: RETRIEVE EVERYTHING PUBLISHED WITHIN THIS AREA or LIMIT THE EXISTING SEARCH TO THIS AREA. In Figure 2, the red box drawn around Tasmania shows how this could be implemented.

Interactive maps can be used to define arbitrary areas of interest. This is especially useful for regions that have informal or conceptual names, such as “The Midwest”; “Silicon Valley” (the boundaries of which are not stable); or New England;.... Having prescribed a region, the latitudes and longitudes within the area can be referred to a gazetteer to identify all named places in (or near) the defined area to support a search for each of these places.

Objective 6. Demonstrate the use of *spatial relationships* in searching. Through inspection of geographical coordinates (latitude and longitude), spatial relationships can be calculated in the form:

FIND PLACES WITHIN 50 MILES OF FRESNO, CA

The command, Find all places within 50 mile radius of place X, would be especially useful in borderland areas and for identifying closely related places that are or once were separate, such as Minneapolis and Saint Paul; Charleville-Mézières; Buda and Pest; Detroit, MI-Windsor, Ontario. If you don’t find enough on a small town, there might be usable aggregate data for the county.

Objective 7. Demonstrate extending searches from the catalog to other geo-referenced datasets, such as museum collections and digital cultural heritage collections. If a bibliographic search can be extended to non-library resources available through, for example, the ECAI Metadata Clearinghouse, then the scope for contextual data and analytical comparison is greatly enhanced. A user who searches for books about Native American mounds could then also see maps of those mounds and their distribution, including links to websites with more information about them. See Attachment One for some examples of digital datasets that could be combined with search results from a library catalog.

Objective 8. Document and disseminate the project findings, through an informative project website, technical reports, papers and poster sessions at conferences, published papers, and openly web-accessible prototypes.

Additional possibilities. The research proposed does not exhaust the possibilities. The intent is to show that significant improvements on the capability of online library catalogs are possible. There are additional intriguing possibilities which we will explore if resources allow. In particular, the use of maps of political jurisdictions (countries, states, counties) is well established, but with georeferencing and gazetteers, other specialized (digital) maps could also be used, e.g. using a language atlas as surrogate for ethnic areas (Cultural areas do not always fit political boundaries, notably in Africa.); climate zone maps (What material do you have about agriculture in high altitude places?); or historical atlases (What is there on architecture in the [former] French province of *Quercy*?).

A disclaimer: The use of computers for geographical analysis is a complex specialty known as GIS

(Geographical Information Systems). This project is concerned with searching for geo-referenced data in catalogs and gazetteers and will use tools developed in GIS, primarily for map displays. We believe that we have the expertise to do what is proposed and have some familiarity with GIS, but we do not claim to be GIS specialists, and we do not regard this project as being “a GIS project.”