

NAMING THE LANDSCAPE: BUILDING THE CONNECTICUT DIGITAL GAZETTEER*

By Scott R. McEathron, Patrick McGlamery, Dong-Guk Shin
Ben Smith, Yuan Su

Abstract: This article describes the background and processes of the initial work of building the *Connecticut Digital Gazetteer*. The background provides the context of the development of digital gazetteers as a national research initiative within the distributed geolibraries paradigm. The actual technical processes of importing and parsing the data are described and analyzed. To convert point data into data with a geographic “footprint,” spatial extents were identified and added using Geographic Information Systems software and the Land Use / Land Cover (LU/LC) data. The processes of metadata creation for the Connecticut History Online project are then analyzed as well as the actual referencing of this data for use in the cartographic interface. A basic analysis of geographic metadata within the larger context of efforts to merge controlled vocabularies is presented. The article concludes by presenting an outlook on future developments in research on digital gazetteers.

Introduction

Ubi or Where is...?

It is doubtful there is a map librarian alive (or dead for that matter) who has not pondered a scrap of a map and wondered... where... is it from, does it go, or does it show?

When one puts in hours on the university library’s general reference desk, the prevailing question is... *ubi*? For example, where can I find a poetry criticism, a citation, and an article on the national energy policy... *ubi*?

It is reasonable to go to a map when we ask, “*ubi*.” But when we go to the library catalog and ask “where,” we must do it with a textual search. However, a map illustrates topological relationships; relationships of ‘near-ness,’ ‘in-ness’ and ‘far-ness.’ Text does not. Text does not indicate that Chicago and China are worlds apart; in fact, the text-based catalog puts them rather close to each other, alphabetically, on the screen. A map interface would indicate where Chicago and China are in relation to each other, and would indicate that Zinjiang is in China...

* Paper presented at the 67th IFLA General Conference, Boston USA, August, 16-25 2001. Section on Geography and Map Libraries.

even though it is quite far from it in the catalog. A map interface, a graphical, cartographic interface has the potential of enhancing access to the catalog for many users.

In order to use a cartographic interface to textual place-names in an on-line catalog we must provide the names with spatial awareness. The names must be in a coordinate system. In order to make use of topological relationships; the place-names must have areal geometry from which spatial neighbors can be derived. A digital gazetteer must be built which geographically enables the on-line public access catalog (OPAC). Such a tool has precedence in the Digital Library and Digital mapping research initiatives of the past decade.

In this paper we aim to provide a basic background of building the *Connecticut Digital Gazetteer* by briefly describing some of the primary research initiatives such as distributed geolibraries and the development of digital gazetteers. We will then describe the processes and framework for building the *Connecticut Digital Gazetteer*. We will conclude by providing a vision of digital gazetteers.

Distributed GeoLibraries

In June, 1998 the National Academy of Science's Mapping Science Committee convened a workshop to explore:

- a vision for geospatial data dissemination and access in 2010
- comparisons of different efforts in digital library research, clearinghouse development and other data distribution
- suggestions of short and long term research needs
- identification of policy and institutional issues

“The Mapping Science Committee serves as a focus for external advice to federal agencies on scientific and technical matters related to spatial data handling and analysis. The purpose of the committee is to provide advice on the development of a robust national spatial data infrastructure for making informed decisions at all levels of government and throughout society in general.”¹

The National Academy of Sciences' Mapping Science Committee (MSC) has been directing research initiatives since its 1993 report *Toward a Coordinated Spatial*

¹ Mapping Science Committee.
http://www4.nas.edu/cger/besr.nsf/web/mapping_science?OpenDocument

*Data Infrastructure for the Nation*² set the stage for the National Spatial Data Infrastructure (NSDI). The report established the Federal Geographic Data Committee (FGDC) and the FGDC's Spatial Metadata Content Standards. Subsequent MSC reports have addressed specific components of the NSDI, including: partnerships in *Promoting the National Spatial Data Infrastructure Through Partnerships*,³ in 1994, basic data types in *A Data Foundation for the National Spatial Data Infrastructure*⁴ in 1995, and future trends in *The Future of Spatial Data and Society*⁵ in 1997 and most recently *Distributed Geolibraries; Spatial Information Resources*⁶ in 1999.

The June 1998 workshop met to build a national vision of a GeoLibrary. The participants asked:

- What will it take to build distributed geolibraries?
- What economic incentives can be put in place such that stakeholders in all sectors of the community (business, education, and government) can and will participate?
- What arrangements need to be put in place in the form of institutions, regulations, standards, protocols, committees, and so forth?
- What research needs to be done to address problems and issues for which no methods or solutions currently exist?
- What data sets need to be constructed, and what mechanisms might be used?
- What software needs to be written, and who is likely to write it?

These are the questions that are driving current national and international research agendas in the mapping sciences. The digital gazetteer has emerged as part of that research agenda.

Digital Gazetteers

A preliminary study of digital gazetteers was held at the Smithsonian Institution in Washington, D.C. in October 1999. The goals of the two-day workshop were to

² Mapping Science Committee. *Toward a Coordinated Spatial Data Infrastructure for the Nation*. Washington, D.C.: National Academy Press, 1993.

³ Mapping Science Committee. *Promoting the National Spatial Data Infrastructure Through Partnerships*. Washington, D.C.: National Academy Press, 1994.

⁴ Mapping Science Committee. *A Data Foundation for the National Spatial Data Infrastructure*. Washington, D.C.: National Academy Press, 1995.

⁵ Mapping Science Committee. *The Future of Spatial Data and Society*. Washington, D.C.: National Academy Press, 1997.

⁶ Mapping Science Committee. *Distributed Geolibraries; Spatial Information Resources*. Washington, D.C.: National Academy Press, 1999.

(1) develop an understanding of the potential of indirect spatial referencing of information resources through geographic names and (2) to identify the research and policy issues associated with the development of digital gazetteer information exchange.

The development of interchangeable sets of geographic name data (gazetteers) and interoperable gazetteer services could result in a major improvement in seamless access to and use of a wide variety of information resources through indirect geospatial referencing. The two-day workshop was convened (1) to develop an understanding of the potential of indirect spatial referencing of information resources through geographic names and (2) to identify the research and policy issues associated with the development of digital gazetteer information exchange. -- The vision developed at the workshop is the Digital Earth metaphor for organizing, visualizing, accessing, and communicating information provides a powerful enabling framework for marshalling the resources needed to understand and mediate environmental and social phenomena.⁷

Definition and scope of digital gazetteers

With three key attributes, a gazetteer supports several functions of an information retrieval system:

- It answers the "Where is" question (for example, "Where is Storrs?") by showing the location on a map.
- It translates between geographic names and locations so that a user of the information system can find collection objects through matching the footprint of a geographic name to the footprints of the collection objects. For example, "What aerial photographs cover parts of Tolland County?"
- It allows a user to locate particular types of geographic features in a designated area. For example, the user can draw a box around an area on a map and find the schools, hospitals, lakes, or rivers in the area.

Beyond these basics, a digital gazetteer needs to support:

- the representation of variant names
- information about the names such as authority, etymology, source, and time span for the use of the names geographic footprints (coordinates representing

⁷ Hill, Linda. "Digital Gazetteer Information Exchange (DGIE) Final Report of Workshop Held October 12-14," 1999
http://alexandria.sdc.ucsb.edu/~lhil/dgie/DGIE_website/DGIE%20final%20report.htm

point, bounding box, polygonal, and linear features) information about footprints such as accuracy, measure method, source, and time span

- descriptive text
- data such as population and elevation, and
- relationships between named places (e.g., an 'Is Part Of' relation between a city and a county).

Building the Connecticut Digital Gazetteer

Technical Considerations

The first step was importing Geographic Names Information System (GNIS) data from the Alexandria Digital Library (ADL)⁸ and then storing the data in a local Oracle database. The ADL gazetteer, designed by University of California, Santa Barbara, is a relational database schema⁹ based on the ADL Gazetteer Content Standard¹⁰ and implemented on an Informix relational database management system (RDBMS). Although Informix and Oracle are both relational databases, to import data from Informix database to the Oracle database directly is not convenient. We used the Extensible Markup Language (XML) data format as middleware for importing the GNIS thesaurus (described below). XML has become the standard for information interchange due to its flexibility and simplicity. The designation of both Document Type Definitions (DTD) for the XML documents and the relational database schema are based on ADL Gazetteer Content Standard, however, there were some differences between the relational database schema and the DTD for XML data. So to import XML data into our Oracle database, our solution proceeded in three steps:

1. We parsed the XML documents to machine-recognizable elements. We used Xerces-1_0_3 which is a XML parser from Apache¹¹ to do the parsing.
2. We built the database schema for our local Oracle system. Since the designation of the relational database schema ADL gazetteer use is based on the gazetteer content standard, and the schema provides portability to all RDBMS, we borrowed the schema and use it on our Oracle system.
3. We observed the difference between the DTD and database schema, and developed a mapping method that maps the XML data into our Oracle database.

⁸ Alexandria digital library, <http://www.alexandria.ucsb.edu/adl.html>, 1998

⁹ ADL gazetteer relational database schema, http://alexandria.sdc.ucsb.edu/~zheng/alex-imp/new_gaz/, 1999

¹⁰ ADL Gazetteer Content Standard, http://alexandria.sdc.ucsb.edu/gazetteer/gaz_content_standard.html, 2000

¹¹ apache XML parser, <http://xml.apache.org/>, 1999

To extend our GNIS database, we added data from *Connecticut Place Names* (CPN) gazetteer of historical place names. Here we also proceeded in three steps:

1. We scanned the whole book and used optical character recognition software (OCR) to recognize its characters and saved the data in text files.
2. We built a parser with Java code which parsed the text file and extracted three types of data for each record of place, they are 'place name,' 'geographic type of the place,' and 'citations' for each record.
3. Finally we joined the CHS data to the existing records in GNIS database and stored them in related fields.

Identifying Spatial Extents of Populated Places

The Geographic Name Information System (GNIS) has identified many types of geographic features as points on a map, including Populated Places (ppl) and streams. It is possible to find the spatial extents of many of those named features using Geographic Information Systems (GIS) software and geospatial data.

Of the many themes of geospatial data available, the one that has been extremely useful in identifying the spatial extents has been the Land Use / Land Cover (LU/LC) data classified from the LANDSAT Thematic Mapper satellite imagery. This classification was done by the University of Connecticut's Laboratory for Earth Resource Information Systems (LERIS) within the Department of Natural Resources Management and Engineering. The resulting data set is a 1998 LU/LC grid of 30 by 30-meter cells covering the entire state of Connecticut. The data have been processed into 28 classes, including residential, commercial, farmlands, forests, etc.

Using this data we can identify cells which we consider urban centers or built up areas by aggregating the following four classes of data into polygons: urban residential, medium residential, rural residential and tree and turf complex. Most of these polygons can then be considered populated places. Then, using capabilities available in GIS and database management utilities, such as spatial overlays and text joins, polygons can be selected or excluded and assigned the information from the GNIS ppl points. Thus, we are transferring the named point to an inferred polygon.

Problems may occur when the urban centers are not separated from each other in the LU/LC data due to connected development or urban sprawl. In this case a single polygon may contain several ppl. Points. Clipping the polygon by known administrative boundaries can sometimes solve this problem. In Connecticut, these are town boundaries. Other times a named ppl will occur where there is no

identifiable development in the LU/LC data. In this case the ppl may be a historical place, or a cartographic “locator” with not enough development to be identified. We have found that a significant number of these Populated Places are road intersections that may have been more populous or otherwise significant in the past.

Identifying streams is an easier task since the Connecticut Department of Environmental Protection (ConnDEP) has processed the hydrographic data set (vectors from USGS 1:24,000 Digital Line Graphs) from the GNIS. Similar processes of text joins and spatial overlays can be used to merge the GNIS identification numbers to these streams and to place the GNIS names into unnamed or misnamed streams. Problems arise where there are gaps in the streams due to lakes or marshes that have not been identified by the ConnDEP as part of the stream.

Some problems arise when the urban centers are not separated from each other in the LULC data due to connected development, or urban sprawl. In this case a single polygon may contain several ppl points. Clipping the polygon by known administrative boundaries can sometimes solve this problem. In Connecticut these are town boundaries. Other times a named PPL will occur where there is no identifiable development in the LULC data. In this case the PPL may be a historical place, or a cartographic “locator” with not enough development to be identified. We have found that a significant number of these PPLs are “Corners” which may have been more populous in the past.

Connecticut History Online

In 1999 the Thomas J. Dodd Research Center at the University of Connecticut, the Connecticut Historical Society and the Mystic Seaport Museum partnered in the Connecticut History Online project.¹² The project is partially funded by a National Leadership Grant from the Institute of Museum and Library Services (IMLS). The current collaborative grant will contain a total of about 14,000 historical images from the partners’ historical photographic collections when it finishes in 2002. These images can be searched or browsed in variety of ways, including by keyword, subject, creator, title and date. Geographical sites may be searched using a digital gazetteer developed by the University of Connecticut’s Map and Geographic Information Center. Descriptions of the images are included in detailed cataloging records.

¹² Connecticut History Online. <http://www.lib.uconn.edu/cho/index.htm>

The selection criterion included knowing the geographic place of the photograph. Selectors were asked to select only those photographs whose place was identified. The catalogers, working with the MARC format, noted the place in locally adapted 651 field. The 651 field provides the most 'hooks' or coded geographic data for a digital gazetteer, particularly the 2nd indicator '7' and the sub-field '2' 'Source of heading or term' enables the Digital Gazetteer to process most efficiently.

The catalogers were instructed to use one of four thesauri, *gnis*, the Geographic Name Information System, *tiger*, the Topologically Integrated Geographically Encoded and Referenced, *chs*, the Connecticut Historical Society's *Connecticut Place-names* and a *local* thesaurus in the sub-field 2 of the 651 field. The *gnis* includes all of the words on the USGS topographic map series and is in a standard format. It is available online from the US Geological Survey <http://mapping.usgs.gov/www/gnis/gnisform.html> and from the "Getty Thesaurus of Geographic Names" http://shiva.pub.getty.edu/tgn_browser/. The *chs* is a scholarly, comprehensive study of historical place names, it is in print format only, but will be scanned and entered into the Connecticut Digital Gazetteer. The *tiger* thesaurus is a database of street name and auto address location. The Official TIGER look-up site is <http://www.census.gov/cgi-bin/gazetteer> and MapQuest; a value-added site is at <http://www.mapquest.com/>. Items in the *local* thesaurus are compiled by the catalogers and added to the Connecticut Digital Gazetteer as needed.

Mapping the CHO database

At present the CHO database numbers in excess of 4,500 records. These records list over 6,500 instances of place names. There seemed to be two methods for providing spatial access to the CHO Online Public Access Catalog (OPAC). One, an 'ambiguous search' would search for place-name strings against the 651 of the OPAC. This method may enable the user to search for the place-name 'Mansfield' and all of the named features in 'Mansfield'. This method was determined to be inefficient. The other method, a 'specific search' provided for pre-processing the names in the 651 field. The 6,750± instances of names represent only 627 different names. For example, there are close to 500 instances of the place-name 'Hartford.' By structuring a specific search for Hartford and more efficient use of the Connecticut Digital Gazetteer, the OPAC and the HTML interface is possible.

An SQL process is performed on the CHO database in the Endeavor system. The query process pulls each instance of a 651 field. After the initial SQL process that extracts the place-names from the CHO database as a table, a series of processes are performed. A table for each thesaurus is created for each category. Second,

each instance of a place-name is counted and the count entered in a field. Finally a URL is constructed from the place-name and the category as a 'scripted search.' These tables are then geo-referenced using ArcView; the *gnis*, *local* and the *chs* against the Connecticut Digital Gazetteer, the *tiger* against the TIGER Street file. These coverages are then referenced through an ArcIMS project.

Accessing CHO through a Spatial Interface

We are using ArcIMS on an NT server to link to the CHO OPAC. When the user determines their geographic choice, a 'scripted' query is sent to the OPAC. The script is in the form:

<http://cho.uconn.edu/cgi-bin/Pwebrecon.cgi?DB=local&SAB1=Hartford&BOOL1=as+a+phrase&FLD1=Place+Name+%28651A%29&GRP1=NOT+with+next+set&SAB2=East+Hartford&BOOL2=as+a+phrase&FLD2=Place+Name+%28651A%29&GRP2=NOT+with+next+set&SAB3=West+Hartford&BOOL3as+a+phrase%28651A%29&CNT=25&HIST=1>. This is a particularly difficult example. It indicates the complications of compound place-names like 'Hartford,' 'East Hartford,' and West Hartford.' This search is for 'Hartford' only. We have constructed the ArcIMS CHO Geolocator is a simple search engine with few tools. It is primarily a cartographic interface, not a GIS application.

The Metadata Framework

Jan Smits asks an important question regarding geo-spatial data access: "can we manage to shift?"¹³ To answer Jan's question, we must look to our recent past. We must ask questions to critically evaluate that past and then be prepared to act. We believe digital gazetteers represent a step in making a shift toward increased geo-spatial data access through *indirect geospatial referencing*, a process of using placenames that are spatially aware to query data.

The efforts that are currently underway within the World's community of libraries and librarians to integrate controlled vocabularies are important developments toward a more effective use of names and subjects in searching. These efforts include the International Federation of Library Association's (IFLA) projects of Universal Bibliographic Control International MARC Core Program, Minimal Level Authority Records (MLAR), and Functional Requirements And Numbering

¹³ Smits, Jan. "Geospatial Data Access: Can We Manage to Shift," *Conference Proceedings, 67th Council and General Conference International Federation of Library Associations and Institutions (IFLA)*, Boston, Massachusetts, August 22, 2001, <http://www.ifla.org/IV/ifla67/papers/165-166e.pdf> (1 October 2001).

of Authority Records (FRANAR). Also, in Europe there is the development of the AUTHOR Project: an effort to network the national name authority files. Our contentions are that while these projects are valuable to libraries – they focus primarily on personal and corporate name authorities and do little to enhance access the geographic data.

Libraries have done a much better job in populating authority databases with personal and corporate names than with subject headings and specifically geographic subject headings. For example, in the United States, there are over five million Library of Congress (LC) form Name Authority records. However, there are only 258,822 total Library of Congress form Subject Authority records – most of these are not geographic. In comparison, the GEOnet Names Server (GNS), which provides access to the National Imagery and Mapping Agency's (NIMA) database of foreign geographic feature names, has over 3.7 million place-names. Also, the Geographic Names Information System (GNIS) has approximately 2 million names for U.S. features. These two databases together total about 5.7 million place-names. The difference in comprehensiveness between these two databases and the total number of LC form of Subject Headings is staggering. If one of the virtues of authority control is “precision” in searching: how can this be achieved when so little of the total existing names are represented in the database?

The structures of library authority control are not keeping pace with current intellectual growth nor information production within the World. Nor is the World library community providing comprehensive authority control for the pre-existing legacy of human information and knowledge, especially geographic information. The structures that are being used to provide authority control, such as AACR2R or MARC21 in the U.S. are not in themselves to blame. The failure of authority control centers on our inability to populate the existing structures with data.

Like authority files, gazetteers have tended to be hierarchical – pointing to a preferred name. The notable exemption is the *Alexandria Digital Library Gazetteer*.¹⁴ This project provides a means of exchanging record formats using a flexible metadata content standard (which is currently being revised), providing a common scheme of categorizing places (feature type codes) and search and retrieval protocols. Key to the success of this distributed approach is how effective local efforts, such as the Connecticut History Online project, are in adding the extra value to the metadata and then sharing it with others.

¹⁴ Hill, L. L., Frew, J., & Zheng, Q. “Geographic names: The implementation of a gazetteer in a georeferenced digital library,” *D-Lib* (January 1999).
<http://www.dlib.org/dlib/january99/hill/01hill.html>

The quality, consistency, and of course, the very presence of geographic coding within any given metadata are key variables for the retrieval of relevant data when using a digital cartographic interface to query a database or multiple databases. For example, the MARC21 map format contains several 'mappable' or geographic elements such as the coded cartographic mathematical data within the 034 field. This field holds the geographic coordinates that define the limits of the map being described. However, there are also other fields that can be mapped. These include the 052 geographic classification code and the 651 subject added entry--geographic name fields. Also, beyond the MARC21 maps format, other subject added entries within the cataloging records for other formats often have geographic coding when they have geographic subdivisions assigned. This raises the possibility of being able to reference this data using a digital cartographic interface as well. In other words, a web based map search interface could potentially serve as a portal to large amounts of geographically coded data including bibliographic data that represents information in many varied formats.

The metadata created for the Connecticut History Online project is an example of geographically coded data that may be queried from both an online catalog or from a map interface. The catalogers for the project used Endeavor's *Imageserver* software to input metadata for each image within the MARC21 format. Within each metadata record, there is at least one 651 subject added entry. The following thesauri are being used to determine the subject added entries: gnis, tiger, chs, ipc, lcsh, local.

Librarians continue to build a virtual distributed international authority file and map and geography librarians need to assess how geographic name authority data may be incorporated. The Connecticut Digital Gazetteer is an example of a database and tool that incorporates many different vocabularies and their corresponding spatial extents. This will enable indirect spatial referencing of data containing coded geographic data. The development of digital gazetteers is an important development toward the realization of the Digital Earth metaphor and will aid and expand the serendipitous discoveries of geographic information.

Patrick McGlamery
Map and Geographic Information Center
Homer Babbidge Library
369 Fairfield Road
University of Connecticut
Storrs, Ct 06269
Patrick.McGlamery@uconn.edu

Scott R. McEathron
Assistant Map and Geography Librarian
University of Illinois at Urbana-Champaign
418 Main Library, MC-522
1408 West Gregory Dr.
Urbana, IL 61801
Phone: (217) 244-189
macmap@ux1.cso.uiuc.edu