Olha Buchel—University of Western Ontario
Linda L. Hill—University of California, Santa Barbara (retired)

Treatment of Georeferencing in Knowledge Organization Systems: North American Contributions to Integrated Georeferencing

Abstract

Recent research projects in North America that have advanced the integration of formal mathematical georeferencing and informal placename georeferencing in knowledge organization systems are described and related to visualization applications.

1. Introduction

Georeferencing using placenames (e.g., Chicago, Ohio River) and place types (e.g., city, river) is used extensively in general conversation, writing, and for knowledge organization because knowing the where and what of places is fundamental to understanding the meaning and relevance of information related to geographic locations. Several studies have shown the extent to which reference to geographic location is used in ordinary discourse and publication. Petras (2004) found that 50% of a test set of five million library catalog records contained one or more place-related subject headings or codes. A company in the business of analyzing text documents to identify geographic references has estimated that at least 70% of the documents they work with contain placenames (MetaCarta Inc., 2005).

For knowledge organization (KO), geographic location is a component of the description and identification of various attributes of information resources (e.g., spatial coverage, geographic aboutness and place of publication) for all types media, whether they are maps, books, articles, data sets, photographs, images, or web sites. Major classification schemes (e.g., Library of Congress, Dewey Decimal), subject heading authorities (e.g., Library of Congress Subject Headings), thesauri (e.g., Getty Thesaurus of Geographic Names, GeoRef Thesaurus), and metadata structures (e.g., MARC) are imbued with geographic terms, codes, and attributes of space. These are types of informal georeferencing that make use of placenames, administrative hierarchies, [1] and place types to

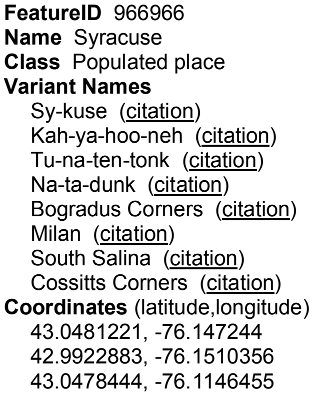
[1] An administrative hierarchy for placenames is based on the partitive relationships among named places based on political units. For example, the placement of Syracuse, NY in an administrative hierarchy within a thesaurus would be "United States – New York – Onondaga County – Syracuse" – or, from the smallest to the largest unit, it can be expressed as "Syracuse is part of Onondaga County is part of New York (state) is part of United States. Another common organizational structure for placename data is a record within a set of records (a catalog) where each record has with fields such as placename, county, state, and country.

designate geographic places. In contrast to informal georeferencing is formal georeferencing, where the location of a place is identified mathematically by use of longitude and latitude coordinates or by use of another global referencing system (e.g., UTM coordinates). Formal georeferencing is fundamental to navigation, cartography, satellite imaging, aerial photography, and the analysis of spatially distributed data. In the last few decades, products and services based on formal georeferencing (e.g., GPS units in our cars, Google Earth, MapQuest, online mapping standards) have expanded greatly due to the emergence and rapid advancements in geospatial technologies. Only recently, within the last decade, has it been demonstrated that bridging between informal and formal georeferencing within KO systems adds powerful benefits in understanding the contents of collections and the relevance and relatedness of information. This paper presents the research and developments that have recently advanced the integration of formal georeferencing into traditional text-based KO, focusing on North American activities.

2. Background

The merits of the use of coordinates in KO are linked to their intrinsic properties of being culturally and language-neutral, cross disciplinary, capable of spatial visualization, and applicable to all types of information resources (Hill, 2006). Coordinates that represent the location of a place can be linked to its placenames in various spellings, languages, scripts, and transliterations, including historical placenames and cultural variants. Table 1 illustrates this with an abbreviated record from the U.S. Geological Survey's gazetteer.

Table 1. Illustration of an entry in a gazetteer showing the preferred name, class, variant names (linked to sources), and coordinates for Syracuse from three USGS topographic maps (abbreviated record from the U.S. Geological Survey's *Geographic Names Information System* (GNIS), http://geonames.usgs.gov.



FeatureID 966966
Name Syracuse
Class Populated place
Variant Names
Sy-kuse (citation)
Kah-ya-hoo-neh (citation)
Tu-na-ten-tonk (citation)
Na-ta-dunk (citation)
Bogradus Corners (citation)
Milan (citation)
South Salina (citation)
Cossitts Corners (citation)
Coordinates (latitude,longitude)
43.0481221, -76.147244
42.9922883, -76.1510356
43.0478444, -76.1146455

Once the location of a place has been expressed in coordinates – even using a simple longitude and latitude point as in Table 1 – it can be situated on a map and its spatial relationships to other places and physical features can be discovered, appreciated, and acted upon. When placenames are associated with information resources, then these resources can also be viewed in a spatial environment and related to resources that are spatially similar in content. Moreover, the use of geospatial coordinates can represent, for example, the location and progression of events such as weather phenomena and migrations (Cahill & Moore, 2006) and vague areas, such as southeastern Illinois, and the resources associated with these events and areas.

Geographic places are listed by name and documented in *gazetteers*, which have typically been structured as dictionaries, encyclopedias, or indexes arranged in alphabetical order, often describing each place in terms of its name and location. An entry in the index of an atlas might look like this: "Nantong, Jiangsu, China Page 23, Grid J2 32.05N 120.51E." An entry in an encyclopedia-type gazetteer might look like this:

> Name of place: **Timbuktu**
> Type of place: **city**
> Location: **Mali**
> Timbuktu (tim-buhk-too), city (1987 pop. 31,925; 1998 pop. 31,973; estimated 2005 pop. 32,460), (cap.) Sixth Region, central Mali, near the Niger River; 16°46'N 03°01° W … an important meeting place for the nomadic people of the Sahara … was founded (11th century) …. Also spelled Tombouctou. (a portion of an entry in the *Columbia Gazetteer of the World Online*, 2005[2])

Other structures for gazetteer data have been used. The authoritative gazetteers of government agencies (e.g., the U.S. Board on Geographic Names and the Geographical Names Board of Canada) use locally designed metadata-like models where there are fields that specify types of data (Table 1). The *Getty Thesaurus of Geographic Names* uses a thesaurus structure. Each gazetteer has been built as a stand-alone reference serving particular purposes with no expectation that the data could be networked to or shared with other applications or integrated with other datasets.

In a review of *Georeferencing: The Geographic Associations of Information* authored by Linda L. Hill, Michael Kennedy notes that geospatial information comes in three forms: (a) maps, (b) numerical coordinates, and (c) text. Geographic information systems, he says, are good at building bridges between (a) and (b), but that all those who are interested in "what is where and why" should be interested in building bridges between (a) & (c) and (b) & (c) (Kenney, 2008). One component to implement these bridges is the development of a formal model for gazetteer data.

[2] Since 2005, the format of the entries in the *Columbia Gazetteer of the World Online* has been changed to include separate data elements for Coordinates and Population.

3. Early Use of Coordinates in Text-based Knowledge Organization Systems

Map librarians were the first to realize the need for fields within the MARC format for coordinate values to document the geospatial boundaries of the maps, aerial photographs, and related materials in their collections. Because of their close association with map users and the geographers who were beginning to develop GIS software, they were aware that the spatial locations associated with the contents of their collections were key parameters of description. The inclusion of coordinates and other parameters of spatial location in MARC in the 1970s preceded major developments in GIS, which began later in the 1980s. The *Anglo American Cataloging Rules* (AACR2) first included a section on coordinates in 1981. These advancements were promoted by map catalogers and have been most consistently used for the description of maps and geospatial data.

Some indexing and abstracting services realized the importance of formal georeferencing early on also. *GeoRef* (American Geological Institute, 2009), the indexing and abstracting service that covers the Earth sciences, started adding coordinates for placenames to its thesaurus and to the metadata for documents in 1977 in order to support a geospatial query capability for its online searching service. The *Getty Thesaurus of Geographic Names* (TGN) (J. Paul Getty Trust - Research Institute, 2009) began adding coordinates in 1987. TGN's scope includes terminology needed to catalog and retrieve information about the visual arts and architecture.

With the emergence of GIS, Nancy Pruett (Pruett, 1986) predicted that digital maps would enhance user tasks, interactions, and retrieval if the contents of collections were geospatially referenced. She foresaw graphical user interfaces for geoscience libraries and information services where a search for maps, journal articles, field trip guidebooks, dissertations, data, and even the names of experts would be carried out by drawing on a computer screen the outline of the area of interest while interacting with an online bibliographic-type database. Ten years later, Ray Larson (Larson, 1996) introduced the concept of geographic information retrieval and explained the advantages of spatial browsing as a method of presenting and querying a variety of georeferenced information using digital maps. This thinking and research, as well as concurrent advances in GIS, prepared the foundation for the next stage of development: building geographically-based digital libraries that demonstrated empirically the advantages of integrating both informal and formal georeferencing into KO and online services, while working out the models and protocols required.

4. Geographically-based Digital Library Projects

4.1 Geo-Referenced Information Network (GRIN)

The first project in North America to design a digital library system that included geospatially enhanced metadata and map-based information retrieval capabilities was called the Geo-Referenced Information Network (GRIN) funded by the Research Libraries Group (RLG). Its goal was to create a library and retrieval system to provide geographically-based access to item-level metadata characterized by geospatial location and then digital access to the actual electronic collection items. The GRIN design included a thesaurus linking placenames to geographic coordinates and graphic displays of the footprints[3] associated with the information resources so that users could see the resources related to their areas of interest ("RLG enters new sphere with geoinformation project," 1989).

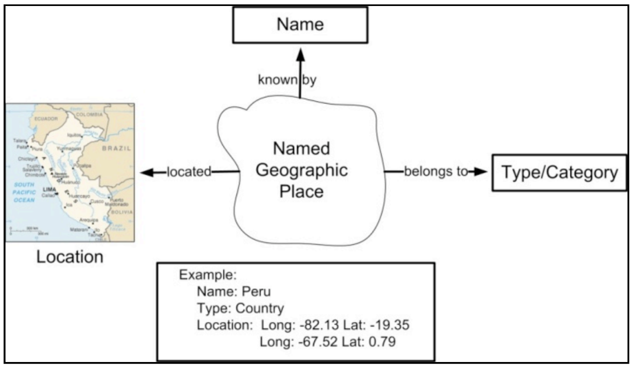
4.2 The Alexandria Digital Library Project and Digital Gazetteers

The first operating prototype of a georeferenced digital library that integrated informal and formal georeferencing was inspired by the GRIN project. The Alexandria Digital Library (ADL) project was developed at the University of California, Santa Barbara (UCSB) as one of the six National Science Foundation digital library projects in the first round of digital library funding, 1994-1998. ADL was designed as a geographically-based digital library (DL) in which the geospatial associations of all types of information resources (e.g., books, articles, maps, remote sensing images, photographs) can be represented by longitude and latitude coordinates and where a gazetteer is integrated as a reference source and to support the translation between placenames, coordinates, and place types (e.g., city, lake, airport) (Figure 1). In ADL, a map-based user interface can be used to display the geographical distribution of resources in a collection, to narrow a search for information to a specific region, and to display the geographic locations of individual resources in the retrieved set. A user can express the geographic location of interest either by placename or by marking an area on the map; that is, either informally or formally. Such a search can be directed to the gazetteer to find out, for example, what “schools” or “lakes” are in an area or to the collections to find resources related to the area as represented either by coordinates or placenames.

The ADL architecture and supporting protocols are based on a distributed system model where collections can reside at distant sites with a shared agreement about methods of generating queries, receiving queries, and returning results. The ADL concept of DL architectures includes the tight integration of *KO resources* (e.g., gazetteers, thesauri, taxonomies) with *collections* and *services*, as presented in a paper presented at the 13th ASIS&T SIG/CR Classification Research Workshop in 2002 (Hill, Buchel, Janée, & Zeng, 2002).

[3] A footprint is a representation of the spatial location or extent of a geographic object or feature represent in terms of a geospatial reference system, such as longitude and latitude coordinates or grid references (Hill, 2006).

Figure 1. Basic components of an entry in a digital gazetteer for a named geographic place: name, location (footprint), and type/category (Hill, 2006, p. 92)



Example:
Name: Peru
Type: Country
Location: Long: -82.13 Lat: -19.35
Long: -67.52 Lat: 0.79

Since digital gazetteers were recognized as key KO components of the ADL Project, a major effort was made to develop a formal data model for gazetteer data and a thesaurus of terms to categorize [4] named places. Using the ADL *Gazetteer Content Standard* (GCS) (Hill, 2004) and the *Feature Type Thesaurus* (FTT) (Hill, 2002), a gazetteer of nearly 6 million entries, with worldwide coverage and assigned categories using the FTT, was created by combining the data from the two U.S. federal gazetteers and other smaller sets of data. This required mapping from dissimilar data structures and local typing schemes. A gazetteer protocol and a thesaurus protocol were created to operate in a networked environment and to support gazetteer and thesaurus query and response services; these protocols do not require that the gazetteer and thesaurus data be in any particular format.

The GCS contains a small set of required elements and an extensive set of optional elements to document aspects such as calendar dates (for names, relationships, footprints, population data, etc.), sources, language, confidence (certainty about the data), authority (e.g., official status of the name), and additional descriptive information. The FTT has six top terms, 210 preferred terms, and 1046 non-preferred terms.

Both the GCS and the FTT have been adopted and adapted for other implementations worldwide. Workshops on gazetteer research and development have been held as a result of the ADL project and the complexities and issues of gazetteer development and implementation have been reported in various publications to further support research in this area and the development of integrated georeferencing in KO (Beaman, Wieczorek, & Blum, 2004; M. K. Buckland & Lancaster, 2004; Crane, 2004; e.g., Hill, 1999; Hill, 2006; Hill, Frew, & Zheng, 1999; Janée, Frew, & Hill, 2004; Kornai & Sundheim, 2003; Networked Knowledge Organization Systems/Services Group, 2002; Smith & Crane, 2001).

[4] The terms *type/typing*, *category/categorize*, and *class/classing* are used interchangeably in this paper.

The importance of gazetteers in KO has been demonstrated and, as more implementations develop, the multiple roles for gazetteers in online information systems are being discovered as well. As a basic reference tool, gazetteers can provide information about a place, such as where it is; what the authorized names for the place are according to various authorities and what other names it has; how it is related to other places; what type of place it is, according to a structured set of categories; and how its names, boundaries, political relationships have changed through time. All of this information supports cataloging and indexing of information with geographic associations. Gazetteers with, perhaps, only name and coordinate information, enable operations such as orienting the map to a named place on Google Earth and MapQuest. For management of placenames by government entities and businesses and for KO implementations, gazetteers are the primary datasets – datasets collaboratively built or with shared access, perhaps, since knowledge of places and their characteristics is most often held locally or collected centrally for particular purposes.

4.3 Other Innovative Georeferenced KO Initiatives

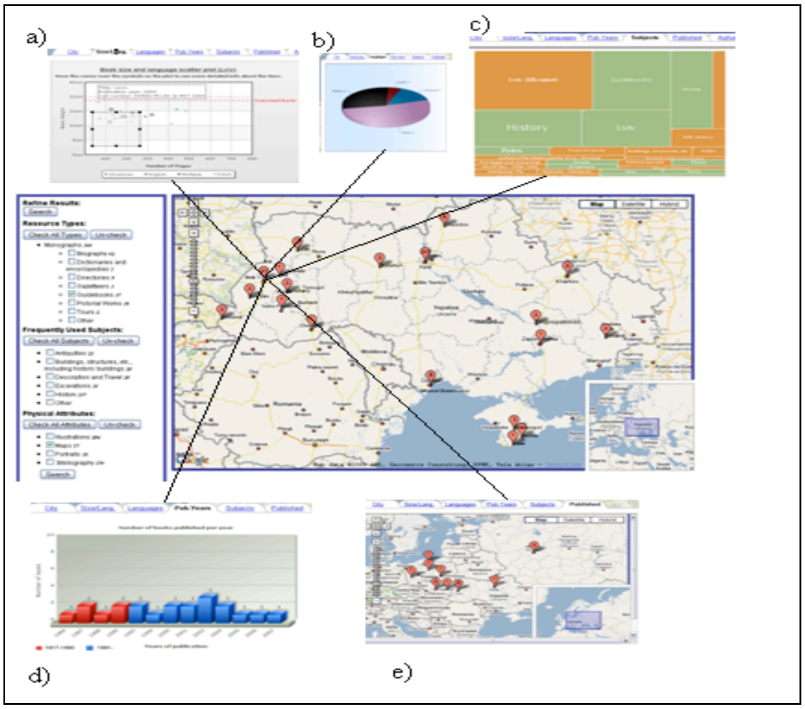
There are operational information systems today that have implemented geospatial referencing for a plethora of information resources beyond maps and aerial imagery, including books, parts of books, learning objects, news articles, genealogical and archival records, historical records, and museum collection metadata. For example, the biodiversity community developed the Darwin Core metadata standard, based on the Dublin Core model, "to facilitate the exchange of information about the geographic occurrence of organisms and the physical existence of biotic specimens in collections" (Taxonomic Data Working Group, 2007). The standard includes a set of georeferencing elements, including coordinate values. The worldwide community that uses the Darwin Core includes natural history museums, zoological and botanical gardens, and germplasm and genetic resource collections.

Significant work has been done for the visualization of library collections on digital maps. Several projects by the Electronic Cultural Atlas Initiative (ECAI) have experimented with the visualization of library collections on digital maps. One ECAI project used a digital map to facilitate searching a collection of 700 MARC records about, or published in, the Cebuano region of the Philippines (M. Buckland et al., 2007). Another project, *Going Places in the Catalog: Improved Geographic Access* (M. K. Buckland, Gey, & Larson, 2002), has experimented with the translation of spatial queries drawn on a map in various graphical forms to text form, and time/space visualizations of library collections using the TimeMap software (Archaeological Computing Laboratory - University of Sydney, 2004), developed in collaboration with ECAI. At the same time, efforts have been made to improve library placename authority records and make them similar to gazetteer records. Since the existing authority records and cataloging practices didn't anticipate this migration to a gazetteer model and to map-based visualization of library collections and information resources, many conceptual and practical issues have to be dealt with in the process.

5. The Future

Google Maps and other online applications are making it surprisingly easy to display data from one or more geospatially-referenced datasets on maps so that the distribution, patterns, and relationships of the data can be seen – or to display a single data point so that its location is shown in the context of its surroundings – or to find the best route between two places. The only requirement is that the places and information resources have coordinate values associated with them, either as recorded in the collection-level or resource-level metadata or because the information systems have placename lookup services that accesses gazetteers to find the coordinates associated with placenames.

Figure 2. Map-based visualization of library collections, where each location is represented by: a) book-size scatter plot; b) language pie chart; c) Kohonen Map of subjects; d) histogram of the years of publication; e) map of places of publication.



Enabling map-based visualization of collection contents adds powerful exploration and discovery interfaces for all types of libraries, archives, data centers, museums, and other managers of knowledge content. Several projects are already underway experimenting with visualizations of resource collections and their contents using digital maps. 4W Vocabulary Mapping project (M. Buckland & Shaw, 2008) visualizes personal biographies as a series of small georeferenced events and links the locations of those events to textual resources (bibliographies, bibliographical dictionaries, catalogs, and encyclopedias). Buchel (2008), as part of her dissertation, develops a prototypical interactive map-based visualization based on a set of MARC records, with links from the geographical locations of the places of publication to dynamic statistical graphics and abstract graphical representations of other attributes from the MARC fields. An example is shown in Figure 2. Here you see the map with icons for sets of books about geographic locations (the map in the center of Figure 2). The linked graphics include a scatter plot of book-size data (Figure 2.a) that allows users to view the distribution of the size of the books published about a particular geographical location; a language pie-chart (Figure 2.b) that visually depicts the languages of books and how many in each language; a Kohonen Map of subjects (Figure 2.c) that shows the distribution of subjects; a histogram of publication years (Figure 2.d) for each location; and another map that shows where the items about the geographic location were published (Figure 2.e).

Pioneering research and development projects in North America, as summarized here, have been important steps in bridging between the georeferencing practices of text-based KO practices and the geospatial practices of GIS. On both sides, the realization is growing that thinking spatially applies to all types of information, to all types of information exploration and use, to all types of knowledge organization.

References

American Geological Institute. (2009). *GeoRef Information Services*. Retrieved March 11, 2009, from http://www.agiweb.org/georef/index.html.

Archaeological Computing Laboratory - University of Sydney. (2004). *TimeMap: Time-based Interactive Mapping*. Retrieved March 20, 2009, from http://www.timemap.net/.

Beaman, R., Wieczorek, J., & Blum, S. (2004). Determining space from place for natural history collections. *D-Lib Magazine, 10*(5). http://www.dlib.org/dlib/may04/beaman/05beaman.html.

Buchel, O. (2008). *How Georeferences in Library Classifications and Bibliographic Attributes in MARC Can Be Used to Crystallize Knowledge about Library Collections (Poster presentation)*. Tenth International ISKO Conference. Montreal, Canada.

Buckland, M., Chen, A., Gey, F. C., Larson, R. R., Mostern, R., & Petras, V. (2007). Geographic search: catalogs, gazetteers, and maps. *College and Research Libraries, 68*(5), 376-387. http://metadata.sims.berkeleyedu/geographicsearch.pdf.

Buckland, M., & Shaw, R. (2008, August 5-8). *4W Vocabulary Mapping Across Diverse Reference Genres.* Paper presented at the Tenth International ISKO Conference, Montreal, Canada. *Proceedings*.

Buckland, M. K., Gey, F. C., & Larson, R. R. (2002). *Going Places in the Catalog: Improved Geographic Access*. Retrieved June 7, 2009, from http://ecai.org/imls2002/.

Buckland, M. K., & Lancaster, L. (2004). Combining place, time, and topic. *D-Lib Magazine, 10*(5). http://www.dlib.org/dlib/may04/buckland/05buckland.html.

Cahill, C., & Moore, S. (2006). Where are we with coordinates? *Documents to the People (DTTP), 34*(4), 37.

Crane, G. (2004). Georeferencing in Historical Collections. *D-Lib Magazine, 10*(5). http://www.dlib.org/dlib/may04/crane/05crane.html.

Hill, L. L. (1999). *Digital Gazetteer Information Exchange (DGIE) Workshop*. Retrieved June 7, 2009, from http://www.alexandria.ucsb.edu/gazetteer/dgie/DGIE_website/DGIE_homepage.htm.

Hill, L. L. (2002). *Feature Type Thesaurus*. Retrieved June 7, 2009, from http://www.alexandria.ucsb.edu/gazetteer/FeatureTypes/FTT_metadata.htm.

Hill, L. L. (2004, 2004-02-26). *Guide to the ADL Gazetteer Content Standard, version 3.2*. Retrieved February 26, 2004, from http://www.alexandria.ucsb.edu/gazetteer/ContentStandard/version3.2/GCS3.2-guide.htm.

Hill, L. L. (2006). *Georeferencing : The Geographic Associations of Information*: MIT Press.

Hill, L. L., Buchel, O., Janée, G., & Zeng, M. L. (2002). Integration of Knowledge Organization Systems into Digital Library Architectures: Position Paper. In J.-E. Mai, C. Beghtol, J. Furner & B. Kwasnik (Eds.), *Advances in Classification Research. Proceedings of the 13th ASIST SIG/CR Workshop on "Reconceptutalizing Classification Research"* (pp. 62-68). Philadelphia, PA.

Hill, L. L., Frew, J., & Zheng, Q. (1999). Geographic names: the implementation of a gazetteer in a georeferenced digital library. *D-Lib Magazine, 5*(1). http://www.dlib.org/dlib/january99/hill/01hill.html.

J. Paul Getty Trust - Research Institute. (2009). *Getty Thesaurus of Geographic Names Online: About the TGN.* Retrieved March 11, 2009, from http://www.getty.edu/research/conducting_research/vocabularies/tgn/about.html.

Janée, G., Frew, J., & Hill, L. L. (2004). Issues in georeferenced digital libraries. *D-Lib Magazine, 10*(5). http://www.dlib.org/dlib/may04/janee/05janee.html.

Kenney, M. (2008). Georeferencing: The Geographic Associations of Information (Review). *The Professional Geographer, 60*(2), 288-289.

Kornai, A., & Sundheim, B. (2003). *Workshop on the Analysis of Geographic References, May 31, 2003, Edmonton, Alberta, as Part of the North American Chapter of the Association for Computational Linguistics and Human Language Technology Conference (NAACL-HLT 2003)*. Retrieved June 7, 2009, from http://people.mokk.bme.hu/~kornai/NAACL/.

Larson, R. (1996). Geographic information retrieval and spatial browsing. In L. C. Smith & M. Gluck (Eds.), *Geographic Information Systems and Libraries: Patrons, Maps, and Spatial Information* (pp. 81-123). Urbana-Champaign: Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign.

MetaCarta Inc. (2005). *MetaCarta. Corporate brochure*. Retrieved August 30, 2005, from http://www.metacarta.com/docs/Corporate_Brochure_06_05.pdf.

Networked Knowledge Organization Systems/Services Group. (2002). *Digital Gazetteers - Integration Into Distributed Digital Library Services*. Retrieved June 7, 2009, from http://nkos.slis.kent.edu/DL02workshop.htm.

Petras, V. (2004). *Statistical Analysis of Geographic and Language Clues in the MARC Record* (Technical report for the "Going Places in the Catalog: Improved Geographical Access" project, supported by the IMLS National Leadership Grant for Libraries, Award LG-02-02-0035-02. ed.): University of California at Berkeley.

Pruett, N. J. (1986). State of the art of geoscience libraries and information services. In E. P. Shelley (Ed.), *Proceedings of the Third International Conference on Geoscience Information, Adelaide, South Australia, June 1-6, 1986* (pp. 15-30). Adelaide: Australian Mineral Foundation.

RLG enters new sphere with geoinformation project. (1989). *The Research Libraries Group News, 19*(spring), 3-9.

Smith, D. A., & Crane, G. (2001). Disambiguating geographic names in a historical digital library. In P. Constantopoulos & I. T. Solvberg (Eds.), *Research and Advanced Technology for Digital Libraries. Proceedings of the 5th European Conference, ECDL 2001, Darmstadt, Germany* (2163 ed., pp. 127-136). Berlin: Springer.

Taxonomic Data Working Group. (2007). *Darwin Core Group - DwC. TDWG Task Group*. Retrieved March 20, 2009, from http://www.tdwg.org/activities/darwincore/.