Digital Gazetteer Information Exchange (DGIE)

Final Report of Workshop Held October 12-14, 1999
Funded by the National Science Foundation, Program 98-121 (Digital Government), Directorate for Computer & Information Science and Engineering.

Larry Brandt, Program Manager
Principal Investigator: Linda L. Hill, University of California, Santa Barbara
Co-Principal Investigator: Michael F. Goodchild, University of California, Santa Barbara

[Please cite as](#)

Table of Contents

Introduction

Summary

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. A 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 workshop involved principal producer and user communities of gazetteer data in the U.S. and other countries, including experts of various academic, library, data center, and private sector groups who have related interests. This report summarizes the workshop proceedings and proposes research and development areas for current emphasis to advance the availability and usefulness of indirect georeferencing through place names.

Vision

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. Growing stores of geospatial information are being collected worldwide, including remotely sensed images of the Earth from space, aerial photographs, data from ground-based and atmospheric monitoring stations, surveys, and digital cartographic products. These resources are *directly* georeferenced by latitude and longitude coordinates locating their *footprints* on the face of the Earth. The resources are also *about* (i.e., include the area of) named geographic features such as cities, parks, and biogeographic regions that may not be specifically noted. Place names for such features are the key components of gazetteers. Digital georeferenced gazetteers link place names to geographic footprints. Current digital library implementations (for example, the Alexandria Digital Library and the Digital Library Project of the University of California, Berkeley) have demonstrated the use of such gazetteers to provide *indirect* georeferencing to geospatial datasets. Through gazetteer translations, a place name can be converted to a geographic footprint that can then be used to find relevant geospatial datasets. But this is only part of the picture. There are vast stores of georeferenced information in library catalogs, bibliographic indexes, and museums that are georeferenced with place names, not latitude and longitude coordinates. The digital georeferenced gazetteer provides the key to merging these data sources with the geospatial data and thus toward the promise of the Digital Earth metaphor for information management.

Digital gazetteers will become a key tool in everyday life and specialized pursuits, linking many properties about a place to the name of the place. They will be embedded in many applications and also available as free-standing services to support discovery and orientation. They will be introduced through early childhood education and adopted by increasingly user-centered geographic systems. Gazetteer creation tools will be widely available, providing templates and knowledge organization tools so that individuals and groups can create gazetteers for special purposes. Official gazetteers providing authorized (and possibly variant) name forms and standard geographic footprints and type categories will be accessible through the Internet through standard protocols. Electronic atlases and geographic information systems (GIS) will present their gazetteer-type information through a gazetteer interface, which will greatly enhance the geospatial information content to be found by place name through gazetteer services. Catalogers of natural history collections, photography collections, library collections, and such will be able to enter place names and find coordinate values to add to their metadata. Local history buffs will create their rich historical gazetteers in a format that can be integrated with all other sources of georeferenced information. Regional planning districts will contribute their local sites to the available gazetteer data. The scope of these developments will be international, integrating national and regional descriptions of local places with the general gazetteer sets composed by the United States federal government and other organizations.

The vision of the future of gazetteers includes having gazetteer spell checkers routinely available in word processing systems. Electronic news accounts will link place names to digital gazetteers and thus be able to display locations of places on maps interactively. Video coverage will link place names from aural text (or captions) to map displays of

locations. Navigation systems will use richer data to support route finding. Voice-activated user interfaces will be able to handle a verbal plea for help "Which way to Cucamonga?"

Temporal views of changing place names, footprint extents, relationships between administrative areas, and the data related to a place (e.g., climate, population) will be supported by temporal ranges for gazetteer data. Named events such as hurricanes, tornadoes, and migrations could be described in "event gazetteers" showing progressive and generalized footprints of the events.

Gazetteers promise to be the key to seamless georeferenced access to the data sources of governments, both national and state. The resources of the U.S. federal government include the massive georeferenced data collections of NASA, NOAA, EPA, USGS, Census, NIMA, and other agencies, and also the contents of the Library of Congress and the libraries and technical information centers of the federal agencies. A user who needs information about a location should be able to describe that location either by coordinates or by place name and discover relevant information from both data collections and library collections.

This future is not so far off. Much of what has been described above could be in place within five years. Supporting technologies to make it happen include high performance systems, gazetteer service protocols, and knowledge organization standards for attribute semantics and category (type) schemes. Urgency will come from those who already know that they need digital gazetteers to meet the needs of their clients: federal and state agencies and information clearinghouses, entrepreneurs and established commercial firms developing georeferenced products, digital libraries, and user communities such as natural history museum curators, librarians, genealogists, local historians, and regional planning groups.

Research communities will develop around the issues of georeferenced information and digital gazetteers spanning computer science, library and information science, social sciences, natural sciences, cognitive science, and more. The National Science Foundation will identify a research agenda and build a research community to move forward quickly on unresolved issues of gazetteer building and use. Standards organizations will develop gazetteer standards that meet the needs not only of geographic information systems but also of information description and management in a broader context. Policy makers will grapple with issues of intellectual property, economic models and access, confidentiality, reliability, authentication, and related issues as they pertain to place name descriptions.

Background

Many types of information have reference to specific places on the Earth's surface. They include:

- reports about the environmental status of regions
- photographs of landscapes
- images of the Earth from space
- census and economic statistics
- museum holdings
- guidebooks
- yellow pages
- maps
- municipal plans
- audiovisual productions, and
- pieces of music.

These (and many more) are examples of information types that are (or can be) georeferenced to a geographic location (National Research Council, 1999). A prevalent form of georeferencing is through place or feature names, frequently found in bibliographic publications, indexes, and catalogs. Those who seek to identify information relevant to their activities often need to do so by reference to a particular spatial location, which they are likely to be able to describe

by a geographic name. Within an information retrieval environment, an example query is "Find all information relevant to fish and wildlife studies about the Cottonwood Creek study area." The user, in this case, would like to find relevant items that are labeled with or contain the phrase "Cottonwood Creek study area" (e.g., reports and papers) and also pieces of data and information that are about the area but that don't specifically mention the place name (e.g., aerial photos and remote-sensing images). This form of indirect geospatial referencing is supported through the use of gazetteers.

A *gazetteer* is a list of geographic names, together with their geographic locations and other descriptive information. A *geographic name* is a proper name for a geographic place or feature, such as *Santa Barbara County*, *Mount Washington*, and *St. Francis Hospital*. Imprecise areas such as *Southern California* can be included. Names such as *Abbeville 30x60 Minute Topographic Quadrangle*, *Grand Fort Tejon Earthquake Epicenter*, and *Habitat of the Red-legged Frog* are also legitimate gazetteer entries because they name identifiable geographic locations. *Geographic identifier* includes proper names for places as well as other types of identifiers: street addresses, postal codes, and *prepositional references* such as "across the Mall" and "5 miles south of the bridge" found in natural language descriptions.

There is remarkable diversity in approaches to the description of geographic places and no standardization beyond authoritative sources for the geographic names themselves. Among the geographic name products that exist are the products of the U.S. Board on Geographic Names, the name authority files of the Library of Congress, the geographic name sets created by indexing and abstracting services, the gazetteer products of other nations and international bodies, and the various sources of spatially-defined geographic names such as digital mapping, GIS datasets, environmental research, and commercial gazetteer products.

A goal of Digital Gazetteer Information Exchange (DGIE) is to enable the interchangeable use all of this data and more while documenting the original source, authority, and accuracy of the data for appropriate use and evaluation. A second goal is to establish the standards and agreements necessary for the interoperability of gazetteer services.

Potential uses of spatially-defined gazetteers can be understood by considering existing projects and websites:

- the online gazetteer services of the U.S. federal government: the Geographic Names Information Service (GNIS) < http://www-nmd.usgs.gov/www/gnis/> and the GEOnet Names Server < http://164.214.2.59/gns/html/index.html>
- the digital library developments at UC Berkeley <http://galaxy.cs.berkeley.edu/> and UC Santa Barbara <http://www.alexandria.ucsb.edu>
- government activities such as the design for NASA's Earth Observing Satellite Data and Information System (EOSDIS)
- the spatial referencing projects of biodiversity data and natural history collections
- commercial facility locators such as those of the Yahoo Yellow Pages
- electronic atlases and online mapping services
- Microsoft's Terraserver < http://terraserver.microsoft.com/>, and
- navigation systems for automobiles.

There is already an awareness of the vast potential of extending spatial referencing to library catalogs and online bibliographic files. One indexing and abstracting service, the AGI's GeoRef (American Geological Institute, 1999), has already implemented spatially-defined geographic names linked to the indexing of their database and the place names in the GeoRef Thesaurus. A result of the National Research Council (NRC) Distributed Geolibraries workshop, June 15-16, 1998, was the identification of gazetteers as a key component of geolibraries (National Research Council, 1999).

The Alexandria Digital Library (ADL) < http://www.alexandria.ucsb.edu> has been engaged in

major gazetteer development since the beginning of the DLI-1 funding period in 1994. A paper describing this development has been published electronically in D-Lib (Hill, Frew, & Zheng, 1999). The ADL Implementation Team combined the two major U.S. federal government gazetteers into one gazetteer containing nearly 6 million entries. As a result of this experience, ADL developed a Gazetteer Content Standard <http://www.alexandria.ucsb.edu/gazetteer/gaz_content_standard.html> and a Feature Type Thesaurus <http://www.alexandria.ucsb.edu/gazetteer/FeatureTypes/index.htm> and has reloaded the U.S. federal gazetteer to this new model. It has added additional gazetteer data pertaining to earthquakes, volcanoes, topographic map quadrangles, and political areas. Where possible, it has added spatial footprints that show the extent of the feature rather than just a representative point location. This gazetteer is one of two major collections in the Alexandria Digital Library and is accessed along with the ADL Catalog to answer both "where is" and the "what's there" type questions.

Related spatial standards work is proceeding in the ISO Technical Committee on Geographic Information (TC211) http://www.iso.ch/meme/TC211.html. These include a proposed standard for Feature Cataloging Methodology and one for Indirect Spatial Referencing. The Open GIS Consortium http://www.opengis.org/ also has activity focused on "Locational Geometry Structures" and "Semantics and Information Communities" that are related to the issues of indirect geographic referencing and gazetteers.

The remainder of this Report describes activities and results of the Digital Gazetteer Exchange Workshop that was held October 12-14, 1999, in Washington DC. There is also a separate narrative report created from the recorded tapes of the presentations and discussions at the Workshop. Other DGIE documents include the Workshop agenda and the list of participants with biographical statements.

Digital Gazetteer Information Exchange Workshop

A workshop was held on October 12-14, 1999. The Digital Government Program of the National Science Foundation provided funding. Other sponsors were the National Geographic Society, Environmental Systems Research Institute (ESRI), Rand McNally, and Carl Stephen Smyth of Microsoft. A reception was held at the National Geographic Society's Explorers Hall on the evening prior to the workshop. The workshop sessions were held at the Smithsonian's Ripley Center.

The goals of the workshop were:

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

Workshop participants were selected through a combination of invitation and open call. Sixty-six people participated in the two-day workshop. The largest group by organizational affiliation was the federal government; the largest group by application area was the biological and environmental group.

	Fed Gov	Academia	State Gov	For Profit	Non-Profit
U.S.	24	17	5	9	4
Non-U.S.	4	2		1	

Application Areas	Number
Biological/Environmental Applications	14
Geospatial Applications	11

Gazetteer Producers	10
Information Science & Library Applications	9
Computer Science	8
Geography	7
Social Sciences	3
Other	4

The following user communities were represented:

- Official government place name authorities
- Entrepreneur product developers
- Geospatial data standards developers
- Information systems researchers and developers
  - Clearinghouses
  - Digital libraries
  - Bibliographic indexes
  - Traditional libraries - map libraries
  - Data centers
  - Data providers
  - Electronic atlases
  - Digital Earth systems
  - Natural history collection information systems
- Information retrieval researchers
- GIS researchers and application developers
- Policy impacts, issues, implications

The workshop agenda consisted of plenary sessions with seven perspective presentations and discussion sessions focusing on three topics:

1. Users of Gazetteer Data and Services: Using and Contributing Gazetteer Data
2. Gazetteer Authority Files: Establishing Authorities and Sharing Data
3. Gazetteer Components of Information Services: Libraries, Clearinghouses, Information Infrastructures

Discussions in four small groups and a summing up and strategy session followed the plenary sessions.

Definition and scope of digital gazetteers

The essential elements of a digital gazetteer entry are:

1. a name
2. a geographic footprint
3. a type or category.

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

1. It answers the "Where is" question (for example, "Where is Santa Barbara?") by showing the location on a map.
2. 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 Santa Barbara County?"

3. 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 volcanoes 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 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 IsPartOf relation between a city and a county).

Especially important information content for gazetteers includes:

1. Identification of the temporal aspects of names, geographic footprints, data, description, and the relationships between places (e.g., change of administrative boundaries)
2. Declaration of the uncertainty associated with footprints and data values
3. Documentation of the source and authority of the names, the footprints, and the data recorded in a gazetteer.

There are a mathematically infinite number of "places" in space when *direct* representation (e.g., with latitude and longitude coordinates) is used to locate them. A nominal set of labels (place names) is used for *indirect* reference and is the primary means used to communicate location in human discourse. These names describe a fuzzy geography, where the perception of a place name's location and extent will differ from person to person (Barr, 1999) and by the context of use. Goodchild uses the example of the interpretation of "Santa Barbara" while in Washington, D.C.; the meaning is understood to be a more general area than the use of "Santa Barbara" while on the outskirts of the city itself. Some systems of naming, for example postal addresses, are well organized for locating places precisely; but even for these there is often a degree of uncertainty based on the extrapolations used to geolocate them. Fuzzy locations are often "good enough" for communication without precision. Spatially defining place names requires solutions for spatially representing fuzzy areas.

Digital gazetteers are the bridge between the vague and the precise in the representation of geographic location; between human cognition and the exactness of scientific representation.

The gazetteer concept has applications beyond the representation of named places on the surface of the Earth. Examples of other applications include gazetteers of other earthly bodies (e.g., the Earth's moon) and gazetteers of fictional worlds (e.g., Tolkien's Middle Earth (Tolkien, 1937)). Usually named places are relatively fixed locations such as physiographic and hydrographic features (e.g., mountains, valleys, lakes, oceans) and administrative entities (e.g., countries, states, counties, cities). Event gazetteers are possible, however, to hold the footprints of such events as hurricanes and floods where the temporal dimension of the footprints represent the geographic location during the life of the event.

The term "gazetteer" is not well understood beyond the groups that construct them (e.g., the U.S. Board of Geographic Names) or use them (e.g., reference services of libraries). The term originated from its use by an English newspaper (a "gazette") for its list of authoritative forms of place names (Oxford University Press, 1971). It is now being revived by increased focus on indirect geographic referencing but, because the term is unfamiliar to most people, the term "gazetteer" has to be repeatedly explained. Some think that the concept should be renamed. Suggestions for a new name include "geolocator."

Research issues:

1. Can we devise a test to determine if a dataset is or is not a gazetteer? Is it simply a matter of including the core attributes or is it a matter of functionality and intent also?

2. Currently footprints are represented in digital gazetteers in two dimensions: x and y (latitude and longitude). What adjustments would have to be made to include the z dimension (vertical; height and depth)?

3. With a larger view, what is the scope of the gazetteer concept beyond geographic locations on the Earth and how does this larger potential scope affect the design of a gazetteer description model and service? For example, the "geolocator" approach can be applied to place and feature names on other planetary bodies (e.g., the Earth's moon). Other potential applications include astronomical locations of objects in space (i.e., where to point the telescope) and anatomical locations within human bodies.

4. How does the concept of the gazetteer, which describes named geographical locations (presumed to be fixed locations), apply to the description of events such as hurricanes? To describe such an event, a temporal sequence of locations and extents can be described; this series of footprints can be aggregated into an overall representation of the total extent of the event.

5. Can the characteristics of geographic names themselves be used in creating gazetteers? Are there patterns of naming that are reliable indicators of category so that they can be used for type assignment? How do these patterns vary by category of place (e.g., lakes versus buildings) and by language, and by culture?

Content of digital gazetteers and representational issues

Place names

Since Adam and Eve started the business of naming everything, we have creatively named most of the features of our universe. Names can be in the vernacular or officially established; any particular place or entity can have a variety of names. Names can be relatively precise (e.g., street addresses) or deliberately vague ("Southern California"). They can represent named regions that exist by virtue of a great variety of relationships (e.g., statistical similarities) or that are official or commercial administrative areas. There are names for natural features and cultural features. The names can indicate the type of place clearly (e.g., *Isla Vista Elementary School)* or the name can be misleading (e.g., *Gilcrease Hills* as the name of a subdivision); not be informative at all about the type of place it is (e.g. *Ekersall)*; or have several possible interpretations (e.g., *Davis Field).*

Commercial names are often a set of names unto themselves, found in "yellow pages" but not in official gazetteers. For example, the Board of Geographic Names has a policy of not including names indicating private ownership. Official names are also carefully monitored to avoid names that include derogatory references.

Place names exist in all languages and in all character sets. They are not unique; that is, the same name can represent multiple locations and entities, even in the same location (e.g., the name of a country can also be the name of the island where it is located). Projects and enterprises can have very different meanings associated with generic names such as the *Western Region*. In current gazetteers, where multiple entries for the same name are present, it is not evident if the multiple entries are about the same entity (with information from different sources) or if there are indeed multiple entities being represented by the same name.

Conventions for the syntax of place names also vary. Library cataloging rules specify how place names for various kinds of subject headings are formatted. There are also rules about how hierarchy is included in names. For example, for one location we use "Paris, Texas" and for another we use "Paris, France." When place names are used within a particular context, the hierarchy is usually indicated only by the context itself. For georeferenced place names, it can be argued that the hierarchy need not be included in the expression of the name at all because inclusion ("is contained by") can be derived from the georeferencing.

Temporal dimension of place name data

Geographic names, their footprints, their relationships to other places, and their associated descriptive elements all change over time. Gazetteers must therefore incorporate temporal ranges for this data and gazetteer services must support both searching and display by temporal attributes.

Since these date ranges are often inexact or estimated, this temporal *uncertainty* must also be represented. Date ranges for geographic names will also need to be extensible to geologic and pre-historic times. Time translations from different time schemes will be needed.

Gazetteer entries themselves also have temporal attributes for creation and modification dates, which need to be represented as part of the administrative information for the record.

Groups working with historical information, such as the Getty Art Institute <http://www.getty.edu/> and the Electronic Cultural Atlas Initiative (ECAI) <http://www.ias.berkeley.edu/ecai/>, will be particularly interested in developing the temporal representation elements of gazetteers and gazetteer services.

Footprint representation

1. *Fuzzy* **footprints:** Since the extent of a geographic feature is often approximate or ill-defined (e.g., Southern California, the Rocky Mountains), rules and methods by which these *fuzzy* boundaries and locations are derived and presented to users are necessary.

2. **Generalized footprints:** Generalized footprints include points that represent the approximate center of a feature (or points derived by some other method) and bounding boxes that represent the maximum latitude and longitude extents of a region. These generalized footprints are more easily available and easier to process in information retrieval systems. It is not well understood when and how such generalizations "suffice" for information representation and retrieval.

3. **Feature extents:** Currently point representation of location prevails in digital gazetteers, often derived as by-products of map production. Points give no indication of the extent of the feature and limited support for the spatial *overlaps* and *contains* evaluations used in geospatial systems to match a query footprint (e.g., the area of interest to the user) to the footprints of items in the collections. Points are used, for example, to disambiguate one stream from another by representing the mouth of the stream or the point that one stream joins another. But the extent of the stream needs to be represented to answer such questions as "What streams exist in the Philadelphia area?" The challenge for digital gazetteer development is how to add the extents of features to existing gazetteers - where are the sources of these footprints? How can they be merged into digital gazetteers?

4. **Multiple footprints:** There can be multiple representations of geographic location for a place or feature: from different sources, for different time periods, of different types (e.g., points, bounding boxes), etc.

5. **Accuracy of footprints:** For appropriate interpretation of footprints, each should be explained in terms of its accuracy and measurement method.

Publishing special purpose gazetteers

Sources of place name descriptions include local planning agency studies, local history projects and surveys, natural history collections, genealogy and personal histories, place name histories from national parks (e.g., Yellowstone), states, regions, etc., atlases, maps, marketing studies, geological and geophysical studies, and so forth. Software, templates and publishing protocols to enable the creation and publishing of this gazetteer data are needed to support the flow of this information into broader gazetteer building efforts and to support the sharing of this type of information in general.

Authority issues

A place can be given many names - in different languages, in variant spellings and syntax, and at different times in its life, for different purposes. Authoritative bodies, such as the U.S. Board on Geographic Names, standardize the form of the names of geographic places for the purpose of clarifying communication and referencing. Newspapers also adhere to similar naming conventions. Gazetteer services must convey the authoritative designations of place name forms clearly while also optionally including variants of the names. The processing needs of gazetteer authoritative bodies must also be considered in designing gazetteer services so that they can receive input from various

gazetteer-publishing efforts for review and action.

Costs and benefits of content enhancement

Potential enhancements to gazetteers and gazetteer services should be judged according to identified benefits for tasks and user communities and the cost of implementation. Priorities for development should be ordered by the greatest benefit for the least time, effort and investment. User communities should drive these decisions.

Change management

A shared environment for gazetteer data and a network of gazetteer services will bring opportunities for synchronizing and announcing/alerting changes in gazetteer information (e.g., the ability to find, or receive notice of, changes made to authoritative gazetteer datasets). Instituting synchronization services and avoiding outdated presentations, however, is a difficult data management problem that needs to be addressed. Metadata tagging of the currency and authority of pieces of gazetteer data will be particularly important in some operations (e.g., emergency management, public safety, military operations). Historical tracking of changes will be important in other operations (e.g., references for historical events).

Quality aspects

Several aspects of gazetteer data quality need to be addressed. One is how to indicate the accuracy of latitude and longitude data. Another is the need to ensure that the reported coordinates agree with the other elements of the description. In general, data quality checks should be built in wherever possible for all data elements.

Multicultural issues

Digital gazetteers and associated toponymic exchange standards must accommodate an extended Roman alphabet and non-Roman writing systems to reflect accurate place name spellings. Selection and implementation of eight-bit (ISO 8859) and 16-bit (ISO 10646/Unicode) encoding present issues of cross-platform compatibility, information storage and processing, and availability of fonts required for text display.

Cultural sensitivity issues arise also. For example, respect for a religious or sacred site might affect the representation of it in a gazetteer.

Privacy and protection issues

Access to footprint references and other information for some locations needs to be limited to authorized groups in some cases. Examples include sensitive archaeological and habitat locations and locations related to military or national defense activities. If any of the information associated with these locations is to be integrated into gazetteers that are generally available, then the content and service structures must be able to protect the sensitive pieces of information from unauthorized use. Generalized footprints can be specifically used when the exact locations cannot be indicated, for example.

Digital gazetteer architecture

Gazetteer data sources can be designed as:

- free-standing gazetteers, such as those of the U.S. Board on Geographic Names and the Alexandria Digital Library, or
- functions of services designed for other purposes, such as GIS services and electronic atlases.

In either case, elements of shareable gazetteer data and services will include:

- a gazetteer content standard that specifies the semantics of the attributes

- community agreement on the values and formats used to represent gazetteer data, and
- gazetteer service protocols.

Services that need to be supported include those that support:

- import and export of gazetteer data
- creation and maintenance of gazetteer data by individuals and organizations for personal, group, and official purposes
- network protocols for search and retrieval of gazetteer information
  - by place name
  - by geographic footprint
  - by feature type (e.g., hydrographic features, lakes, cemeteries)
  - by temporal range for place name, footprint, etc.
  - with spatial matching operators (e.g., overlaps, contains, is contained by, is x distance from)
- registries for gazetteer services or a gazetteer domain for network addressing, and
- gazetteer interfaces for GIS and electronic atlas data sources

User interface designs for gazetteer creation and search and use need to be prototyped and tested for various applications and user communities.

Gazetteer content standard

A content standard defines a set of attributes and a structure for those attributes, covering all of the important aspects of gazetteer description, as discussed below. This provides a general framework for describing geographic places. The intellectual design of such a standard comes first, with community agreement on the set of attributes and their meanings. The standard must have concrete requirements as well as flexibility in expression to meet the needs of diverse applications. The resulting standard is formally defined for implementation (e.g., Universal Modeling Language, XML DTD).

The Alexandria Digital Library has developed and implemented a Gazetteer Content Standard (Alexandria Digital Library, 1999). It provides a basis for further development. It is designed on a metadata model where each record (entry) is self-contained and can be shared with other applications. This follows the practice of library cataloging where the MARC record structure (ANSI-NISO Z39.2) has provided the basis for shared cataloging among libraries at great cost savings to each individual library. Another example is the FGDC Content Standard for Geospatial Data (U.S. Federal Geographic Data Committee, 1998) that provides the basis of the National Spatial Data Infrastructure (NSDI) with over 150 distributed servers of geospatial data. The alternative to the metadata model is the hierarchical thesaurus model such as the *Thesaurus of Geographic Names* created by the Getty Art Institute (Getty Information Institute, 1997). Place name thesauri tend to arrange place names hierarchically on a whole/part basis. For example, *Chicago* is a narrow term of *Cook County*, which is a narrow term of *Illinois*. Some place name thesauri, however, are based on generic hierarchy instead. In such a case, Chicago would be a narrow term of *cities* (or a similar term for the type of place it is). Thesauri are inherently very difficult to merge and share because of their embedded structure and contextual assumptions. However, gazetteer data can be converted from a thesaurus structure to the metadata model for sharing with other applications.

The ADL Gazetteer Content Standard is presented in Appendix A. It has the following characteristics:

1. Small set of required data: Only one name, one footprint, and one type are required for a minimal entry.
2. It is designed to merge information about a place from several sources and to cite the source of each piece of information. It is a requirement that each piece of information be traced to its source.
3. Temporal ranges are applied to names, footprints, and description. Dates for record creation and modification are also included.

4. Flexibility is supported in the choice of category type scheme, relationship types (e.g., IsPartOf, IsCapitalOf), and data types (e.g., population, elevation).

5. Accuracy and measurement method of geographic footprints can be represented. This is an important attribute for any geospatial description but for gazetteers it is particularly important because named places often have fuzzy boundaries (e.g., the Rocky Mountains) and because bounding boxes and points are often used to represent the locations instead of more precise polygons.

6. Links can be added to the gazetteer entry to access other data sources for the place, such as details of an earthquake from the Southern California Earthquake Center linked to an entry for the epicenter of the earthquake in the gazetteer.

Workshop participants suggested the following additions to the ADL Content Standard:

● indication of the slope and aspect of a site

● ability to represent three-dimensional representations - adding the z dimension

● behaviors of feature types (e.g., behaviors of a *dam* include the characteristic that it forms a barrier to navigation)

As with other metadata models, there is an issue about what constitutes the minimal content for a gazetteer database entry. The ADL Content Standard specifies three required elements and also requires attribution for the source of each piece of data and basic metadata creation information (e.g., date of creation). The question is whether this set of requirements is sufficient or does it vary with the application. For example, for what purposes are the temporal dimensions required?

Feature types

Type has proven to be a key attribute for description and searching across distributed collections. Adoption or application of common type schemes across collections enables semantic interoperability and generic searching across collections. Without it, a user has no reliable way to discover the type schemes used to organize the various collections and is reduced to the hit-and-miss technique of free-text searching.

There is no common feature type scheme for place categorization in use today. The two sections of the U.S. federal gazetteers from the U.S. Geological Survey and the National Imagery & Mapping Agency have incompatible type schemes. Types schemes of other countries are also based on local sets of terms and relationships among terms. In an effort to fill this hole, ADL developed a Feature Type Thesaurus based on hierarchical thesaurus design as specified in the ANSI-NISO Z39.19 standard (National Information Standards Organization (U.S.) & American National Standards Institute, 1994). The ADL type scheme has been applied to the two U.S. gazetteers and to other sets added to the ADL Gazetteer. Converting original types to the ADL types has proved to be by far the most difficult part of building the ADL Gazetteer. These conversion problems have been described in detail in two papers: *D-Lib* January 1999 (Hill et al., 1999) & a paper presented at the 1999 ASIS annual conference (Hill & Zheng, 1999).

Shared feature type schemes to categorize individual features for shared gazetteers will not be easy to develop but have the potential to facilitate the sharing of gazetteer data and the use of data from multiple sources. These schemes need to be hierarchical (e.g., "lakes" IsTypeOf "hydrographic features"), rich in term variants (i.e., synonymous terms for the same categories), and extensible to accommodate greater depth in terminology where needed. To be practical, schemes designed to facilitate sharing need to incorporate variant forms of terminology from established feature type schemes so that they can provide mappings between the various schemes.

A basic research issue for type schemes (in general, for all thesauri) is how to interrelate them to one another. That is, how to map:

● terms and meanings from one scheme to another scheme, and

● specific classification schemes for particular applications (e.g., GIS feature classes and EPA wetland classifications) to type schemes designed to be generally applicable (e.g., the ADL Feature Type Thesaurus)

The last case is the *macrothesaurus* to *microthesaurus* problem that has been studied in various ways before: the *macrothesaurus* for shared high-level structure linked to the *microthesaurus* for in-depth extensions of terminology. The users of gazetteers - and the sources of gazetteer data - span general public to specific GIS applications. No one type scheme will satisfy all of these contexts. An architecture of type schemes ("concept architectures") will be needed.

Other research issues with feature type schemes:

1. Computational methods for automated support for feature classification based on the characteristics of place names themselves and the pattern of feature type assignment in existing gazetteer collections.
2. Determination of a minimal set of feature type categories.

Gazetteer services

The goal of having distributed gazetteer servers accessible over the web will be realized only when the protocols to support machine-to-machine query-response and import-export have been developed. Figure 1 illustrates the interfaces required. The *public server* of a *gazetteer site* accepts queries from *clients* and returns *output* (results). This is a generalized model; it can build on similar capabilities already in place, but will need to be specialized for gazetteer services.
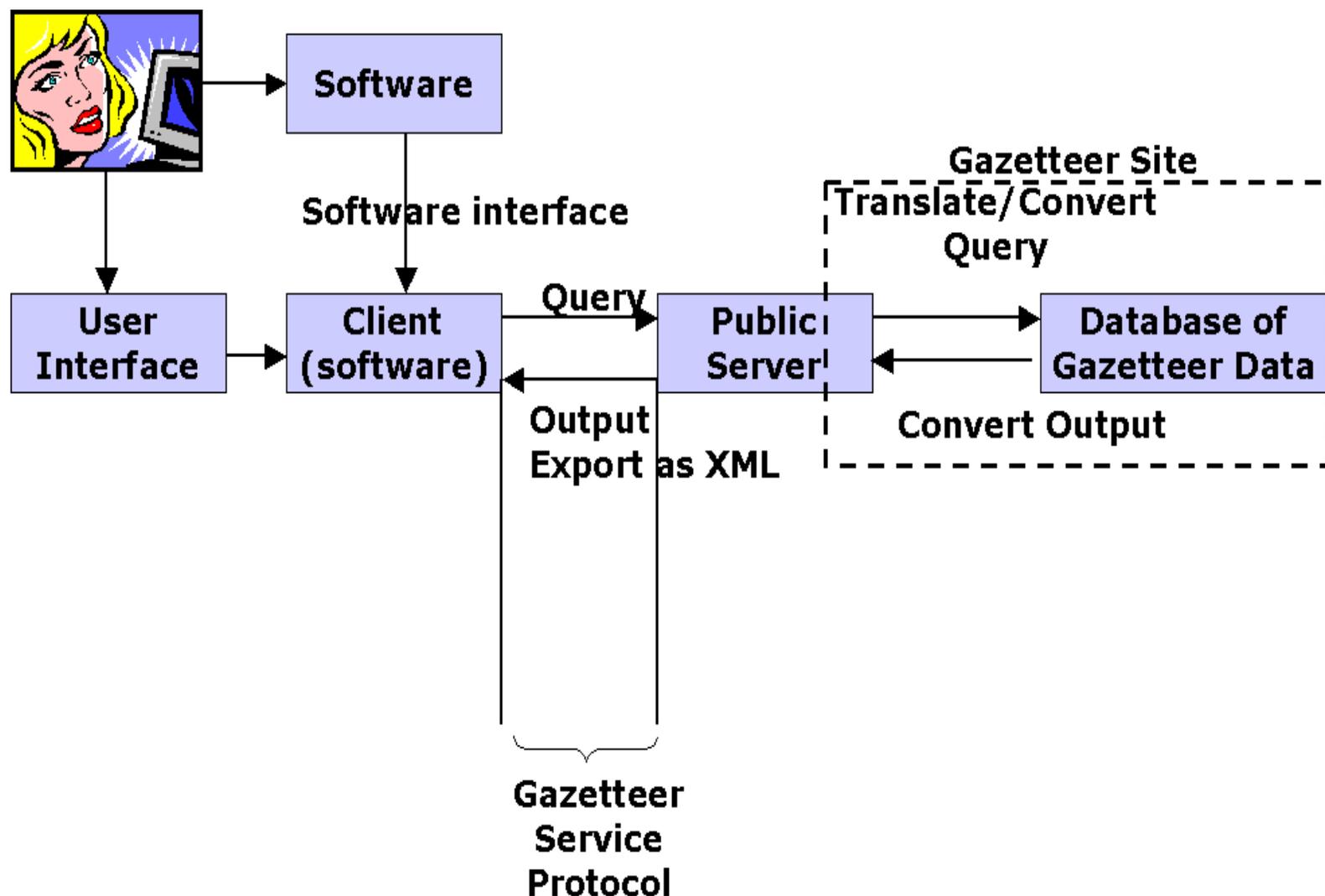

Figure 1. Gazetteer Service Model

Gazetteer query types will include those listed earlier:

- by place name (support for "near matches" and non-Roman characters)
- by geographic footprint ("contains", "overlaps", "distance from" operators)
- by feature type (support for different type schemes; incorporation of thesaurus services)
- by temporal range for place name, footprint, etc.

Users should be able to establish hierarchical preferences for gazetteer servers. For example, the application that calls on gazetteer services might look at a preference list defined by the user of the application. A local gazetteer might be on the top of the user's preference list, containing places of specific interest with local definitions, a combination perhaps of locally created entries and entries downloaded from other gazetteers. These local gazetteers may take precedence over authoritative sources for some purposes.

Gazetteer responses will include both full reports and selected data for gazetteer entries in XML format and according to a gazetteer content standard.

Research issues:

1. Develop ways to allow for natural language to be used in queries and data gathering.
2. Determine the sufficient level, if any, of metric accuracy for footprints necessary to respond to queries.

Standards

There is an obvious need to work on the standards needed to support shareable gazetteer data and gazetteer services. The following specific standards were noted:

- Develop a formal gazetteer content standard similar to the metadata standard for geospatial data, working through the Federal Geographic Data Committee and the ISO TC211 standards group.
- Develop web-based gazetteer service specifications that will translate between geographic names and footprints through the use of an OGC-type process.
- Develop feature type category schemes that harmonize those from the Spatial Data Transfer Standard (SDTS) and other GIS sets, those of the authoritative gazetteer services, and those developed by the library communities.

User communities

A list of user communities and uses of digital gazetteers tends to expand to include everyone and everything since the gazetteer is a basic reference tool that is applicable across the board, especially in an increasingly digital-earth-aware world. Referring to locations by name is a primary representational method. Enabling the translation of these names to geo-locations through digital gazetteers and thus to GIS datasets will be useful to individuals in their everyday pursuits and to community and government activities.

The workshop participants identified the following list of potential user communities in particular:

- emergency response ("What schools, hospitals, etc. are in the area that needs to be evacuated?")
- natural science research ("Where are the collection sites for the flora and fauna of a region?")
- humanities ("Where were the Buddhist temples during a particular time in a particular region?")
- mapping ("Display and label the hydrographic features for this map area." "Where is the Ford dealer in this area?")
- archives ("What historic documents pertain to the upper Ohio River area?")
- environmental applications ("Find the wetlands within 10 miles of the periphery of large urban centers.")
- urban planning ("What information do we have about the buildings in this block?")
- recreation ("What beaches are in this area?" Where is El Capitan?")

- education ("Where is Bosnia?")

Policy issues

The policy issues of gazetteers share the issues of other intellectual property:

- security and confidentiality for sensitive information (particularly spatial locations) of sites that are only appropriate to share with a limited group (e.g., sensitive habitat locations)
- authentication (digital signature)
- intellectual property, rights (constraints), including support of the U.S. position on full and open access to federally created datasets and the implications of international and federal laws and court decisions
- attribution and lineage of the source of the gazetteer information (provenance)
- authority level of the information (e.g., is the data from an authoritative source)
- level of confidence (quality) indicators for the data
- currentness of the data
- appropriate use statements for the data

Other Issues

1. **Frequent changes**: Geographic names, their footprints, and the data and relationships associated with them are constantly changing, leading to maintenance and documentation issues. Ideally gazetteer data will be maintained on a transactional basis and protected with good database maintenance practices.

2. **Merging information from different sources**: Information about a particular gazetteer entry can be obtained from multiple sources. Footprints, description, name forms, geographic data, etc. from various sources are best combined into one record rather than separate records, with attribution for the source of each piece of information. To do this merging with confidence, we need to develop methods of identifying when two sets of information are indeed about the same place. The answer may be to develop a numbering system for places. If, however, we are operating in an environment without an international numbering system for places, then we need to develop algorithms that use available information about the places to determine, with some degree of certainty, that two places are the same. The evidence for such a determination will include the place name, the location, the category of the place, and any additional information provided by the sources.

3. **Computational issues:** The computational issues of gazetteers are largely those of a georeferenced digital library:
   - efficient indexing and searching of large, distributed datasets
   - effective combination of spatial, temporal, quantitative, and textual search processes
   - map-based visualizations of gazetteer entries and sets of gazetteer entries, putting the locations in the spatial context of other GIS data and in relation to other gazetteer entries.

Development paths

This workshop has drawn on the knowledge of various communities to establish the scope and potential of digital gazetteer information exchange, as summarized here in the Final Report and explained in more detail in the Narrative Report of the workshop. These communities gathered here for the first time to discuss this issue. Before this focused discussion, the gazetteer component of information services was perceived solely from the practical viewpoints of the various communities. Those that are responsible for establishing the authoritative names of places at the national and the state level are the most experienced with establishing gazetteers and maintaining them. They, however, have focused primarily on the names and have been satisfied with the use of point locations for the spatial reference. Information services, such as the National Spatial Data Infrastructure (NSDI), have focused on geospatial datasets that

are georeferenced with latitude and longitude coordinates and have not emphasized indirect referencing through place names. Libraries have used place names as subject headings, depending on the authoritative gazetteers for the correct names. Organizations such as the National Geographic and Rand McNally that have traditionally provided place name indexes to geographic atlases are aware of the "translation" function of gazetteers between names and geographic locations. Their new electronic products include versions of digital gazetteers. Very little information science research has considered georeferenced place names and their potential for effective description and retrieval in traditional bibliographic systems. Among digital library research, the Alexandria Digital Library is the only one to emphasize gazetteer development. Among standards development, the only activity in relation to digital gazetteers is the current work of the ISO TC211 Working Group.

It is clear that the digital gazetteer is a component that will play a significant role in the development of a Digital Earth and in linking all types of information to locations. We have established that reality with this workshop. Now we need to take steps to establish gazetteer data and services and the standards needed to build an infrastructure for gazetteer information exchange.

Collaborations and partnerships

The following participants were identified for collaborative efforts to develop digital gazetteers and gazetteer services:

- Federal Geographic Data Committee (FGDC) and the National Spatial Data Infrastructure (NSDI)
- GIS companies
- Companies developing automobile navigation systems
- Open GIS Consortium
- United Nations Group of Experts on Geographical Names (UNGEGN)
- International Congress of Onomastic Sciences (ICOS)
- International Cartographic Association
- Global spatial data community
- Pan American Institute of Geography and History (PAIGH)
- Library of Congress
- International Federation of Library Associations (IFLA)
- International Hydrographic Association
- World Bank
- Place Name Survey of the United States (American Names Society)
- Council of Geographic Names Authorities
- Genealogical societies
- Global Biodiversity Information Facility
- Getty Art Institute
- Networked Knowledge Organization Systems (NKOS)

In addition, the Cooperative Research and Development Agreement (CRADA) vehicle was identified as a process that could be used by the federal government to advance the development of gazetteer products and services.

Next steps

1. Create commercial, academic, and government partnerships to develop gazetteer data and services.
2. Publish papers and monographs and present conference papers on digital gazetteer development and research issues.

3. Advance gazetteer standards through the International Standards Organization (ISO) and the Federal Geographic Data Committee.

4. Generate interest in gazetteer development and gazetteer services through the Open GIS Consortium (OGC).

5. Inventory of existing gazetteers: As a basis for initial development, an inventory of existing gazetteer datasets would be helpful. The datasets that include latitude and longitude coordinates (or other translatable referencing schemes) will be the most easily adapted to online gazetteer services. Gazetteers without coordinate locations will have to be evaluated to see what effort would be needed to georeference them.

6. Identify the initiatives needed to develop or compile interchangeable gazetteer sets with footprints that represent the extent of the feature (as opposed to a point only).

7. Evaluate the uses and users of current gazetteer services and use the results to build user requirements for new gazetteer services and a set of user scenarios that can be used as the basis for metrics to evaluate the effectiveness of new gazetteer services.

8. Study the current and potential interaction modes for gazetteer services, including human and machine interfaces.

9. Identify communities with a need and an interest in building gazetteers for their own purposes and provide the software for gazetteer building and dissemination.

10. Encourage academic research into the cognitive and computational issues of indirect geographic referencing, some of which are identified in this report.

References

Alexandria Digital Library. (1999). *Gazetteer Content Standard*. Available: http://www.alexandria.ucsb.edu/gazetteer/gaz_content_standard.html.

American Geological Institute. (1999). *GeoRef*. Available: http://www.agiweb.org/agi/georef/about.html.

Barr, R. (1999). What's in a name? *GEOEurope, 8*(9), 24-25.

Getty Information Institute. (1997). *Thesaurus of Geographic Names*. Available: http://www.ahip.getty.edu/tgn_browser/.

Hill, L. L., Frew, J., & Zheng, Q. (1999). 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

Hill, L. L., & Zheng, Q. (1999). Indirect geospatial referencing through place names in the digital library: Alexandria Digital Library experience with developing and implementing gazetteers. *Knowledge: Creation, Organization and Use. Proceedings of the 62nd Annual Meeting of the American Society for Information Science, Washington, D.C., Oct. 31-Nov. 4, 1999* (pp. 57-69). Medford, NJ: Information Today.

National Information Standards Organization (U.S.), & American National Standards Institute. (1994). *Guidelines for the construction, format, and management of monolingual thesauri : an American national standard*. Bethesda, Md.: NISO Press.

National Research Council, Panel on Distributed Geolibraries (1999). *Distributed Geolibraries: Spatial Information Resources: Report of a Workshop*. Washington, DC: National Academy Press.

Oxford University Press. (1971). *The compact edition of the Oxford English dictionary : complete text reproduced micrographically*. Oxford Oxfordshire ; New York: Oxford University Press.

Tolkien, J. R. R. (1937). *The hobbit; or, There and back again*. London: G. Allen & Unwin.

U.S. Federal Geographic Data Committee. (1998). *Content Standard for Digital Geospatial Metadata*. Available:

http://fgdc.er.usgs.gov/metadata/contstan.html.

Appendix A. [Alexandria Digital Library Content Standard](#)