Summary Report
Digital Gazetteer Research & Practice Workshop

December 7-9, 2006
Santa Barbara, California

by
Michael Goodchild
Linda L. Hill
Department of Geography
University of California, Santa Barbara

Table of Contents

I. Executive Summary

a. Introduction

The two-and-a-half day workshop focused on the role of digital gazetteers in georeferencing applications, starting with overviews of the state-of-the-art and current activities and leading to a consensus on the opportunities and directions for collaboration and the advancement of a research and practice agenda. The workshop was convened by the National Center for Geographic Information & Analysis (NCGIA) at the University

of California at Santa Barbara and the Redlands Institute. It was sponsored by the National Geospatial-Intelligence Agency. It was held at the Upham Hotel in downtown Santa Barbara, starting with a reception on Thursday evening, December 7th and extending through Friday and Saturday. The organizing committee was composed of the following:

Michael F. Goodchild, University of California, Santa Barbara (Co-Chair)
Linda L. Hill, University of California, Santa Barbara (Co-Chair)
Allen Carroll, National Geographic Society
Tom Elliott, University of North Carolina, Chapel Hill
John R. Frank, MetaCarta Inc.
Jim Frew, University of California, Santa Barbara
Bruce Gittings, University of Edinburgh
Jordan Hastings, University of California, Santa Barbara
Greg Janée, University of California, Santa Barbara
Christopher B. Jones, University of Cardiff
Naicong Li, Redlands Institute
David M. Mark, University of Buffalo
John Wilson, University of Southern California

b. Theme of the Workshop

The workshop had three focus areas:

1. Components of gazetteer services

The three core elements of gazetteers – placenames, place categories, and geospatial locations – support the translation between informal georeferencing using placenames ("Santa Barbara") and place categories ("city") and the formal georeferencing of mathematical schemes (e.g., longitude and latitude coordinate systems). These elements plus explicit relationships between named geographic places and the identification of time frames for places and their characteristics are the fundamental components of digital gazetteers. Within the context of gazetteer services - such as support for enterprise georeferencing systems, geoparsing of text to derive spatial locations, navigation services, and support for geographic information retrieval (GIR) - the complexities of each of these components challenge the collection and use of gazetteer data.

2. Georeferencing as a process

Georeferencing by naming and categorizing natural and human-made geographic features is universal. The practice is highly influenced by individual strategies, local conventions, and requirements of particular applications. This session explored studies that ferret out the nature of the motivations and practices of place naming and categorizing in individual, cultural, historical, information management, scientific, and business contexts and how they inform the construction and use of gazetteers and gazetteer services.

3. Interoperable gazetteer services

Gazetteer data exists in many independent sources often dissimilar in construction and content, including:

- Gazetteers of official toponymic authorities
- Local, formally published, or special purpose gazetteers
- Indexes accompanying atlases
- Place identifier tables accompanying GIS datasets
- Placename authority files used for cataloging and indexing
- Historical printed gazetteers and encyclopedias
- Online sources such as Wikipedia

This session explored the requirements of gazetteer protocols and services to support interoperable access to and use of these distributed sources.

c. Purpose and Structure of the Workshop

The purpose of the workshop was to bring together a wide range of researchers and practitioners who are working with the definition, modeling, and application of gazetteers and gazetteer services to present and discuss current activities and short- and long-term research and development directions. This was achieved by limiting the size of the group and choosing a setting that encouraged active participation in the sessions and interactions before and after the sessions. The organizational structure included three chaired sessions focusing on the themes of the workshop, with two presentations each and appointed discussants who responded to the presentations. Each session included open discussion time. Part of the second day was devoted to breakout discussion groups on specific topics. Participants were randomly assigned to the breakout groups. Reports from these groups to the whole assembly of workshop participants provided another opportunity for the exchange of ideas and the identification of priorities.

The list of participants and their biographies and position papers were posted on the NCGIA website along with the meeting schedule and this summary report (www.ncgia.ucsb.edu). Participants were encouraged to review the final report from the NSF-funded *Digital Gazetteer Information Exchange (DGIE*) workshop, held in 1999; this report was posted with the workshop information.

d. Participation in the workshop

Researchers and practitioners known to the organizers as actively engaged in gazetteers and gazetteer services were invited to the workshop. In addition, an open *Call for Participation* was distributed through a number of multidisciplinary and international email distribution lists. Applicants were asked to submit a two-page statement of interest in the topics of the workshop and a two-page biographical statement. All applications were reviewed by the co-chairs with the assistance of Jordan Hastings. A total of 68 persons were included in our deliberations. Of these, 3 were from sponsoring organizations, 13 were members of the organizing committee, 23 were invited, and 28 were applicants. Eighteen of the applications were not accepted, partly because of space limitations, and 7 of the invitees were unable to attend because of previous commitments. We ended up with 43 in attendance at least part of the time. Seventeen of

these were from non-US locations, including Chile, Germany, Canada, and the United Kingdom. Over half of the attendees were affiliated with universities (27); 8 with for-profit organizations, 7 with government agencies, and 1 with a not-for-profit organization. Areas of interest included cognitive psychology, geographic information analysis, gazetteer building and maintenance, geoparsing, historical scholarship, georeferencing of enterprise information content, mobile navigation services, emergency services, cartographic publishing, georeferenced access to distributed information resources, multilingual information services, geographic ontologies, and what one participant called "placename intelligence."

e. Schedule of Workshop Activities

Thursday 12/7

5:30 pm - Welcome Reception, sponsored by go2 Systems, Inc. (Lee Hancock, President)

6:00 - 7:30 Presentations

Mike Goodchild & Linda Hill (UCSB): welcome and introductions
Jordan Henk (University of Redlands)
Beth Driver and Randy Flynn (NGA)
Chris Rewerts (US Army Corps of Engineers, Engineer Research and Development Center -Construction Engineering Research Laboratory)
Lee Hancock (go2 Systems, Inc.)
Jim Frew (UCSB): Keynote presentation

7:30 - Dinner on your own

Friday 12/8

8:00 - 8:30 Continental Breakfast, Coffee and Tea

8:30 - 9:00 Workshop organization and goals (Mike Goodchild & Linda Hill)

SESSION 1: COMPONENTS OF GAZETTEER SERVICES
(Session chair: John Frank, Metacarta, Cambridge, Massachusetts)

9:00 - 9:30 Allen Carroll (National Geographic)
9:30 - 10:00 Discussion
10:00 - 10:30 BREAK
10:30 - 11:00 Bruce Gittings (University of Edinburgh)
11:00 - 11:30 Discussion
11:30 - 12:00 Response/comments by two discussants: Beth Driver (NGA) and Ray Larson (University of California, Berkeley)
12:00 - 12:30 General discussion and summation
12:30 - 2:00 Lunch (provided)

SESSION 2: GEOREFERENCING AS A PROCESS
(Session chair: David Bodenhamer, The Polis Center at Indiana University-Purdue University, Indianapolis)

2:00 - 2:30 David Mark (State University of New York at Buffalo)
2:30 - 3:00 Discussion

3:00 - 3:30 BREAK
3:30 - 4:00 Chris Jones (University of Cardiff)
4:00 - 4:30 Discussion
4:30 - 5:00 Response/comments by two discussants: Mike Dobson (Telemapics) and May Yuan (University of Oklahoma)
5:00 - 5:30 General discussion and summation
5:30 - 7:00 Demonstrations and refreshments
Dinner on your own

Saturday 12/9
8:00 - 8:30 Continental Breakfast, Coffee and Tea
8:30 - 9:00 Plan for the day (Mike Goodchild & Linda Hill)

SESSION 3: INTEROPERABLE GAZETTEER SERVICES
(Session chair: James Reid, EDINA, University of Edinburgh)
9:00 - 9:30 Greg Janée (University of California, Santa Barbara)
9:30 - 10:00 Discussion
10:00 - 10:30 BREAK
10:30 - 11:00 Ruth Mostern (University of California, Merced)
11:00 - 11:30 Discussion
11:30 - 12:00 Response/comments by two discussants: Paul Ell (Queen's University, Belfast) and Tom Elliott (University of North Carolina)
12:00 - 12:30 General discussion and summation
12:30 - 1:30 Lunch (provided; can be taken into breakouts)
1:30 - 2:30 Breakout Session I
2:30 - 3:00 Reports from breakout groups
3:00 - 3:30 BREAK
3:30 - 4:30 Breakout Session II
4:30 - 5:00 Reports from breakout groups
5:00 - 6:00 Concluding discussion and future directions
7:15 - Conference Dinner at Opals, 1325 State Street, (805) 966-9676

Sunday 12/10
Morning Hike (optional; strenuous 3.5 hour hike planned)
8:00 - 9:30 Continental Breakfast, Coffee and Tea

f. Outcomes of the Workshop

Some outcomes are not immediately obvious, as shown by past workshop experience, because the connections made and the information content of the presentations and discussions has influence beyond the tangible outcomes. Tangible outcomes, however, include this final report, a proposed special issue of the *International Journal on Geographic Information Science (IJGIS)* with contributions from workshop participants, the posting of workshop presentations on the NCGIA website, and the posting of extensive notes taken by a participant, Tom Elliott, on the web at http://www.ncgia.ucsb.edu/projects/nga/ncgia.html.

g. Outline of the Report

This report consists of six sections: the Executive Summary; links to and summaries of the presentations and summaries of the discussions; reports of the breakout sessions; the resulting research and practice agenda; a list of workshop participants; and acknowledgements. The discussion and breakout sessions summaries depend highly on the notes taken by Tom Elliott, the summaries given by the session chairs and the reporters from the breakout sessions. The summaries highlight the main points and are not meant to be a full account of what was said.

II. Presentations and Discussions

Session 1: Components of Gazetteer Services

John Frank (MetaCarta and session chair), as introduction to the session, highlighted the tension between different origins of gazetteers, such as "expert-created" gazetteers on one hand and "wikipedia-type" gazetteers on the other and "stuff" created by people who didn't know that it would be used as a source of gazetteer information.

Allen Carroll

National Geographic Society: *A Case-Study-In-Process: How a Media Organization Tackles the Georeferencing Challenge/Opportunity*

Allen started off with a graphic example of the political pitfalls awaiting those who put placenames on maps in politically and culturally contested areas. The example was labeling the body of water on a map of the Middle East as the "Persian Gulf (Arabian Gulf)" which ignited some very angry responses from those who fervently disagreed with that decision. He followed with a description of the current efforts of the National Geographic Society (NGS) to establish an overall georeferencing strategy so that all of their divisions are using common methods of georeferencing and thus providing a single mode of geographic access to their digital archives and new content. He notes that the key assets of the NGS include their brand name, content, multiple media presentations, cartography, and their ability to tell stories about our world that captivate the attention of readers and viewers. They operate with a goal of becoming "the best source of curated, authoritative information about the world, organized and accessed geographically."

Georeferencing challenges for archival content include little or no consistency in naming conventions and poor or generalized location information. For new content, the NGS is deploying GPS-enabled cameras and video equipment and is incorporating georeferencing (place-names, addresses, descriptive text) into editorial and archiving workflows.

A key component of the new approach to georeferencing is an enterprise gazetteer that will be based on their place-names database for cartography. It is being built as a hierarchical structure based on standards and practices of existing online gazetteer services, fine tuned to their special requirements. The Thesaurus Master software of Data Harmony will be used to merge the gazetteer content from various divisions and

from other sources. NGS is likely to use the MetaCarta text parsing tools to facilitate the georeferencing of text-based content and media metadata and the Red Hen Systems software for geotagging, viewing and cataloging new visual media content. At a later phase, the plan is to fully integrate the gazetteer with web-enabled cartography. The currently operating NGS MapMachine, produced in partnership with ESRI, will serve as an alternate means of access to NGS content.

A specific goal is to "use web-based and mobile mapping applications to tell geography-based stories," such as narratives to accompany map-guided walks through tourist venues. The driving motivations for spatially enabling National Geographic content are (1) to make their content more visible, accessible, and versatile of NGS staff; (2) to enable NGS to better serve new markets: 3-D globe apps, GIS, mobile/GPS, etc.; (3) to form content distribution relationships that will diffuse geographic knowledge—and generate revenue; (4) to integrate multimedia content with cartography in innovative ways; (5) to create a gazetteer that, together with the content and cartography, is of value as a business asset; (6) to provide access to authoritative information about the world from a trusted brand (quality over quantity); and (7) to extend their 91-year tradition of cartography and spatial storytelling into new realms.

Questions/Comments for Allen Carroll

There were comments about authority, trust, and attribution. Comments about annotating the provenance/pedigree of data; how systems track and transmit the origin of the data; if the ancestry of a particular piece of data can be reconstructed; and that trust is derived from transparency. NGS reacts to and uses "complaint letters" that come in, but they don't track these after they make the decision or publicly credit the individuals. What makes a name authoritative?

GNIS tracks the lineage of their gazetteer data. Half of the data comes from paper maps and these names and locations can be tracked back to the surveyor. The other half came from state-level authorities through arrangements with the states. GNIS also has recently implemented a contributor interface; geonames.ucgs.gov.

NGA captures and tracks lineage, analysis, and decision-making in placename research and also tracks temporality of the gazetteer data but does not currently expose this information to users. Such information is available on a case-by-case basis.

Temporality was brought up as an aspect of authority. Time range is a dimension along with placename and spatial location used to disambiguate places.

Bruce Gittings

University of Edinburgh: *Gazetteer Services: A Scottish Perspective*

Bruce started off by defining four types of gazetteers:

- Short-form gazetteers that are often a list of place-names together with their locations expressed with a spatial referencing system; includes the place-name indexes of atlases
- Long-form (or descriptive) gazetteers that may contain lengthy textual descriptions of places and associated maps and photographs
- Thematic gazetteers that are lists of places and their locations by theme, such as fishing ports, nuclear power stations or historic buildings.
- Address gazetteers that contain georeferenced address lists

His talk focused on long-form gazetteers. Within the Scottish context, building such gazetteers is hampered by regularly changing administrative geographies, multiple (nearby) places of the same name, Gaelic place-names with no agreed versions, and the absence of any definitive place-name gazetteer or naming authority. On the other hand, there are many historical maps, detailed country-wide mapping since the 1860s, and a hardcopy Scottish place-names survey from the 1980s. Motivations for building Scottish gazetteer services include place-name linking between heritage projects and supporting efficient government operations and linking numerous online services that provide local navigation and their mapping, imagery, and geographically-referenced texts. In Scotland, many independent resources built with public money could also be linked together through place-names, such as local and central government information resources, the National Archives, the National Library of Scotland, and the Scottish Cultural Resource Archive Network (SCRAN).

A current Scottish project is to build a National Address Gazetteer for Scotland that will be aggregated from local address gazetteers created and maintained by local councils and property tax assessors (diagrams of the infrastructure shown). This address gazetteer is based on property records, to which individual persons can be connected, and will be used by a wide range of government services, such as property assessment, land use planning, legal title registration, census information, mail delivery, refuse pickup, tax collection, student records, police and fire department support, and ambulance services. The data will not be available to the private sector. There are real civil liberties implications to this project since distributed information about individuals will be aggregated as never before. As a national gazetteer, it is limited to only street addresses and post-towns; that is, it does not include non-addressable features (e.g., natural features) and it does not include urban and rural districts, which are so important in defining communities.

A richer collection of information about places can be found in descriptive gazetteers and this information is potentially more comprehensive to the public than maps. The *Gazetteer for Scotland* (GfS) aims to be a definitive and detailed resource. Creating it began in 1995 and currently contains around 13,500 entries with both contemporary and historical text, photographs, and maps; also links to associated people, events, families/clans, video and sounds, and bibliographic references and dates are included as possible. Online access to the GfS is approaching 1 million hits per week. The

website is designed for the general public (examples shown). The service is supported by an Oracle database and by pre-generated pages to facilitate query responses.

Historical descriptive information about places is very rich in details and the subtleties of history and ephemeral knowledge. We don't have good tools to interrogate these resources or make inferences from them and the difficulty of doing so is compounded by the confusions of place naming (e.g., three "Newbiggings" within a few miles of each other in Angus). As part of the GfS project, they attempted to semi-automatically link to *Groome's Ordnance Gazetteer* (1885) through place-name matching but had limited success. Contributing to the difficulty were the number of settlements, country houses, etc. that have disappeared; historical places not listed in the contemporary gazetteer; the changing importance of estates, laird's houses, etc; new developments; spelling changes and inconsistencies; old Gaelic renderings; and inadequate description to distinguish similarly named places.

Bruce concluded by re-emphasizing the enormous value of descriptive gazetteers which contain historical data to lend authority. Such gazetteers can be the glue that joins up other services as well as provide definitive names, statistics, and possibly even definitive descriptions. For Scotland, there is a definite need for a place-name authority and for free access to definitive gazetteer services.

Session 1 Responses

Beth Driver

Functions of Names

- Organize the world & the information in it
  - Finding the information we want
  - Designating specific places/things: uniqueness within an understood or implied universe
  - Defining what things/places are
- Asserting dominion
  - Over the things/places we name
  - Over people who use other names
- Revealing/asserting identity or belongingness—or not
- Revealing attributes of speaker or interlocutor—or both
  - Identity
  - "Reference universe"
  - Deference/respect for interlocutor
  - Deference/respect for place names
  - Formality or informality of conversation

Sets of Names: includes collected names, as in a gazetteer, and used names, as in speech or text

- Collections often have unanticipated value
- Attributes of a *set* may have value
  - What kinds of places we name and why
  - How we set/recognize boundaries

        - Patterns in creation/selection of names, within and across cultures
        - Indicators of change and its direction
        - What makes a set of names "consistent"? Why would we care?
        - Who enforces use of names in practice? How? Why?
    - Are there natural 'fracture lines' for segregating sets of place names?
    - What are implications of wanting to find/filter/process sets of names for how we design/populate gazetteers?

Do Names Require New Services?

- Represent (orthography, transliteration, pronunciation, translation, alternative forms …)
- Store & retrieve
- Discover
- Link to related names
- Spatial reference (qualitative & metric)
- Recognition and capture
- Protect
- Determine quality for individual name and for sets of names
- Presentations and transformations
- Duplicate detection

Ray Larson (University of California, Berkeley)

- Authorities: the Library of Congress as an example (authorities.loc.gov ...)
    - They track variants and sources/attribution
    - Reusing this resource, because it's made available to others
- Reaction: yes, more authorities, more integration of resources across multiple agencies and sources
    - Sometimes linkage won't be perfect, but if adequate, it can be helpful

Session 1 Discussion

Linda Hill

- We've mostly talked about place names, but gazetteers are more than that:
    - Classification of features (types)
        - Not many shared schemes
        - Culturally determined
        - Ruth Mostern: Gazetteers will have very particular kinds of feature type lists ... this is problematic
            - different hierarchies
            - different languages
            - how to deal with this in a way that allows local/domain-specific approaches to persist, but then how to merge/harmonize/federate?
    - Role of IDs for features

        - Do these help us (really) sort out the ambiguity of places and their names?
        - Point made about the problem of permanence of IDs
        - Point made about that this is an important role for authorities
    - If gazetteers are "information tools embedded in information systems, primarily for the purpose of information retrieval,"
        - What kind of footprints belong "in" gazetteers?
            - The more detailed the footprints, the more expensive and difficult to deal with them in the IR context/systems
            - What can you do with generalized footprints? Is that good enough?

Mike Goodchild

- Pronunciation as an issue
    - Computer audio can play a very useful role, but we're not using it
    - Tom Elliott (Pleiades): I wonder whether we should encode pronunciation for our placenames? Or, rather, give our users the option to do so?
        - Whose pronunciation?
    - Randy Flynn (NGA): recognizes the urgent need for further research into these issues, both with respect to names and gazetteer services
        - Relating the spoken placename to other components of place data

Others

- Collaborating with other resources
    - Don't merge it all into one big database - insurmountable resource problems
        - There will always be myriad databases
        - How do you establish linkages and connectivity amongst many thousands of such datasets?
        - How do you evaluate the reliability and relevance of the data you retrieve from interacting with them?
    - Automatic grouping is very difficult (MetaCarta sees this problem)
        - Semi-automated ways? but doesn't work when there's a single place need
        - Users want to reach out to a single bit of data in a single local gazetteer, but they don't know where/what it is nor how to ask that question that way

- RISE, part of INSPIRE process (European spatial infrastructure): data harmonization
    - Purpose-driven gazetteers
    - Use cases are glaring omission from most of this discussion
        - Very hard to identify
    - Grassroots vs. top-down approaches
        - Encourage grassroots within a framework that says harmonization is an essential component of the process

- Krzysztof Janowicz: Is there a formal theory of identity of place?
    - To what degree does "a place" persist, despite time-wise changes in names and geometries and dominions?

Session 2: Georeferencing as a Process

Session chair: David Bodenhamer, The Polis Center at Indiana University-Purdue University, Indianapolis

David Mark

NCGIA-Buffalo: *How Gazetteer Work: Cultural and Linguistic Influences on the Georeferencing Process*

David began by noting that *georeferencing* is a cognitive process and a computational process. Georeferencing supports information retrieval for human users, both the general public and the experts/professionals, and georeferencing is a service used by other computer applications such as by the Semantic Web. The question is "in what ways (if any!) should knowledge of the cognitive process inform the design of computational solutions?" And also, "how different are the requirements for gazetteers in different contexts – such as, in digital libraries, in multilingual information environments, and in regard to sources and users?"

David's main research for the last four years has been on cultural and linguistic differences in geospatial conceptualization and referencing. He is one of the principal investigators for an NSF-funded Ethnophysiography Project that involves an investigation into the ways in which the Yindjibarndi people of Australia and the Navajo people of the western United States describe and communicate about the landscapes that they live in. The project includes five inter-related topics:

- "Geographic Categories": common nouns or non-phrases that refer to *kinds* of geographic things
- Toponyms: proper names for individual geographic features
- Topophilia: emotional bonds between people, place, and landscape
- Indigenous mapping
- Indigenous Geographic Knowledge Systems: for example, traditional stories that incorporate landscape features. Before written language and graphic maps, geographic information was often stored and transmitted in *stories*, which often also included origin stories or moral codes. Place names, and the places themselves, formed "retrieval keys" for the information. (Reference: *Wisdom Sits in Places—Landscape and Language among the Western Apache*, by Keith H. Basso)

The two main methods of fieldwork for this project are (1) *field interviews"* in which the actual language used while out in the landscape is collected and (2) "photo response" in which photographs of landscapes are shown to study participants and the way they talk about them is recorded. The goal of the project is to understand the meanings of words used by "general" speakers of languages to refer to landforms, rather than the use of *scientific vocabularies* for landform types.

David introduced the *semiotic triangle* as a way to consider the complexities of place description within the context of a gazetteer model where the core elements of description are placename, category, and footprint. The semiotic triangle identifies the three corners of a triangle to represent the *concept*, the *instance* of the concept, and the *symbol* representing the concept. For example, there is the concept of a cat, the instance of a particular cat, and the name applied to the concept (or name of the individual cat). Assigning the core elements of a gazetteer entry to the corners of the triangle is a way to represent the complexities and interrelationships of the elements. For example, there is a *concept* of a placename and there is an *instance* of a placename; there is a *concept* of category and there is an *instance* of a category within a set of categories. Relationships between placenames and instances can be made at the *concept* level of meaning and between the *instance* level of meaning. Cultural effects apply to all three corners of the triangle and also to the inferences that can be made. For example, delimitation of the footprint extent for an instance may be influenced by the associated category of the place and that delimitation estimate itself may be culturally sensitive.

For the Navajo, many (most?) geographic features have at least two different proper names: one for traditional (origin) stories and one for every-day use; at least some of the sacred names are used only in the winter!

He gave examples of feature categorical descriptions that are *single words* in some language but which have no single word translation in English:

- A canyon wall receiving sunlight
- A spot of level ground in the mountains, surrounded by ridges
- A type of hollow in a sandhill, used as camping place, especially in cold weather
- An island of land completely surrounded by one or more younger lava flows (e.g., “kipuka” in Hawai’ian)
- An island of grassland left unburnt after a surround wildfire (e.g., “nyirirr” in Walmajarri, an Australian language)

He used the concepts of “lake” and “river” to demonstrate the pitfalls of searching by category across multi-lingual, multi-cultural information resources. In English, standing-water bodies are put into the rough categories of “lake,” “pond,” and “lagoon.” The French have a similar set of terms for categories: “lac,” “étang,” and “lagune.” But because the characteristics used as the basis for these categories differ, these terms are not equivalent. For example, some standing-water bodies that are called “lakes” or “lagoons” are considered to be “étang” (usually translated as “pond” or “pool”) by the French. For the English category “rivers,” the French differentiate between “rivers that flow into the sea – fleuve” and “smaller rivers – rivières”.

The multi-lingual challenge is exemplified by the fact that there are about 5,000 languages in the world that have 1,000 or more speakers. There may be about 100 geographic terms per language, giving 500,000 terms that need to be defined and implemented.

There has been very little work with human subjects on the delimitation of the footprints of geographic features. One study was done in Santa Barbara to collect notions of the boundaries of "downtown Santa Barbara." (Montello, D. R., Goodchild, M.F., Gottsegen, J., and Fohl, P., 2003. Where's Downtown?: Behavioral Methods for Determining Referents of Vague Spatial Queries. *Spatial Cognition & Computation*, 3(2&3), 185-204.) David showed examples of the results of the study: boundaries for downtown Santa Barbara with 100% confidence and boundaries with only 50% confidence.

A related topic is the conception of *landforms*. A claim (definition) of the term is that it refers to a part of the Earth's surface that is identified by shape, occupying a finite region, which has some degree of perceptual or functional coherence of form. The shape of the landform is inherited from the pattern of elevations of the Earth's crust with the boundaries of the landform. Thus, the land surface shape and the landform boundary are mutually dependent. Cross-cultural comparisons of landform definitions have found conceptual differences, which may mean that different taxonomic hierarchies for landforms may be needed for each language or dialect or culture. If a goal of the geographic aspects of the Semantic Web and the Internet is to provide online access to geographic information about any area of the world to speakers of any language, we might have to record feature types separately for each language. However, a better approach would be to develop procedures for feature delimitation and classification that could accept a formalized version of any landform definition, and then extract and classify instances from digital elevation data.

David describes work done toward a strategy for definition-dependent extraction of *topographic eminences* using digital elevation data—"eminence" used as a super-ordinate term for hills, mountains, buttes, mesas, and other such landforms that stand above their immediate neighborhoods. Once eminences are extracted, they can be classified based on properties such as:

- morphographic character: shape, size, position, orientation
- spatial relationships: proximity, prominence, topology
- visual signature: viewshed statistics, angle of depression/elevation, visual prominence

Future work includes attempting to distinguish eminence types named and used in English; attempting to relate Navajo eminence types; extension to other regions and languages; and extend the methodology to other landforms such as canyons and valleys.

Chris Jones

University of Cardiff: *Acquisition of Knowledge of Vernacular and Vague Place Names*
With Paul Clough (University of Sheffield), Hideo Joho (University of Glasgow), and Ross Purves (University of Zurich)

Gazetteers used in online services to translated place names into geospatial locations (e.g., timetables, routing instructions, yellow pages, general web search) tend to reflect an administrative geography. When users of these services enter vernacular place names and names of landmarks instead to indicate a place of interest, the service cannot make a

match; that is, the query fails because of the limitation of the gazetteer. There is a need to acquire knowledge of vernacular names and the location/boundaries of the associated places and add this information to gazetteers for such purposes. Without this component, gazetteers are incomplete and inadequate.

Place name knowledge can be found in places other than formal gazetteers:

- maps and terrain models, which have place annotations that associate names with valleys, mountains, ranges, peaks, etc; difficult to derive the extent of these places
- people through interviews and questionnaires; traditional methods are inefficient, but web methods offer great potential
- text documents; detailed descriptions are difficult to interpret automatically, but associations between places in text may be useful.

Exploiting the Web for place name knowledge is another promising approach. Techniques for Web mining have been developed that derive the spatial extent of vague places in terms of the places that lie within them and Web questionnaires have been developed to elicit personal knowledge of vague places and the use of vague spatial language references to places.

Web mining techniques look for documents that reference precise places located within places with vague boundaries and for place names frequently associated with a target named place. For the latter condition, the assumption is made that places frequently associated with the target place have a higher chance of being inside the boundaries of the target. The steps of the procedure are

- submit web search engine queries referring to a target place
- parse resulting highest ranking web pages for occurrence of place names
- geocode ("ground") place names with coordinates
- create geometric model (surface model) and extract an approximate boundary for the target place.

An example of this method is deriving the boundary of "the Cotswolds". A surface plot of the locations of associated place names shows a concentration around the map location of the Cotswolds. Another example, based on the "Highlands of Scotland," is a surface model showing the density of (a) unique associated places, (b) occurrences of the name of each of these places, and (c) number of documents that mention each place. A third example, based on "Mittelland," compares boundaries derived from human interpretations of extent to the density surface model from Web mining results.

Details of the technique involve formulating appropriate Web queries, methods and rules for geoparsing query results, and methods of geocoding. There are three patterns of Web queries:

- region only; e.g., "Rocky Mountains" – retrieves all documents mentioning the name
- region + concept; e.g., "hotels in Cotswolds" – tends to retrieve directory pages listing places associated with the target place

- region and lexical pattern (trigger phrase); e.g., "Midwest towns such as …" and "in the south of France" – reduces the number of relevant documents retrieved but can work well for those documents; problem of not enough places for statistical analysis

The region + concept technique produces the highest number of co-associated places in top ranking documents.

Geoparsing of the Web query results uses named entity recognition (NER) methods to identify names and gazetteers to recognize the subset of place names. In addition, rule-based strategies are used to distinguish between geographic and non-geographic statements; many place names occur in organization names and in people's names. For example, a <forename> <placename> pattern indicates a person's name.

Since many places have the same name, geocoding techniques need to determine which of the possible named places is being referenced in the documents. The more sophisticated approach is to search for co-occurrence of parent and neighboring places that establish uniqueness. A cruder approach is to assume that the references are to the most commonly occurring instance.

Queries were submitted to Google, via an API, of the form "find hotels in <place>". Initially, precise target places were used (English counties) and the results were compared to the known exact boundaries. Subsequently, less precise target places were used and the results were evaluated qualitatively. Examples of the results were shown. For Devon (county), the distribution of associated places and analysis of the density of places in the Devon area resulted in an estimation of the boundaries that enclosed all but a few small protrusions; the estimated boundary extended beyond the actual boundary to some extent (note: some places were wrongly geocoded). Additional examples for "Mid Wales" and "Cotswolds" were shown.

A new project, funded by the Ordnance Survey and starting in 2007, will investigate the use of Web questionnaires to elicit the personal knowledge from individuals. For the research, people will be asked:

- to give the names of places they live; these names will be associated with the person's (georeferenced) postcode
- to draw the extent of places with vague boundaries
- to label Google Earth with place names
- to give natural language descriptions of vague places; e.g., what places are inside, what are the bounds of the place, and relationships to other places.

Plans for future work include:

- improving Web mining methods
    - thresholding of surfaces needs to be automated
    - quality of geoparsing+geocoding needs to be improved
    - techniques for problem areas that have no human settlements; use queries that target named topographic features

- applications of Web questionnaire methods
    - evaluate different methods of elicitation and geometric modeling
    - use as "ground truth" for web mining
- investigating the use of Web questionnaires for obtaining knowledge of vague spatial language; e.g., west, near, between.
- use of terrain models to extract topographic features
- TRIPOD – European project for image search engines; interpret and generate geographical descriptions of archive images and images for location aware cameras

Session 2 Responses

Mike Dobson (Telematics)
From the point of view of gazetteers as components of local search, how should a gazetteer function? His points:

- We don't always account for the mixtures of semantic patterns that occur geographically.
- An application of geographically-aware searching is targeting ad revenue.
- According to Word Tracker, there are no geographic names in the top 1000 terms searched on the Internet daily; does this have some meaning for gazetteers?
- Frequency of use of geographic words may be a useful comparison to make between competing websites.
- Geographical errors in Web searching are often overlooked because these pale in comparison to failures on other axes; e.g., try a search for "fish tacos in zip code 92653.
- Search companies are buying gazetteer companies.
- What is the definition of neighborhood names and how do you find and exploit them?
- The most important names for marketers (and researchers?) are functional, non-administrative names, but only re-insurers, title insurers, and real estate companies know them.

May Yuan (University of Oklahoma)

- Characteristics of placenames
    - In comparison to personal names, personal names have meaning and are well defined in most cases.
    - Placenames are more complicated. They change through time and there can be more than one name for a place; the characteristics of the place can also change. How do we know that two descriptions of place are about the same place? There are also linguistic difficulties.
    - What constitutes place descriptions? Geographic knowledge is implicit in cultural, historical and locative contexts and is embedded in placenames. The same can be said for the temporal and spatial components of place description. Placename are composite constructs. We don't just want to know about placenames and their associated locations, we also want to know about the stories behind them.

- When we build gazetteers, we need to include for each place its names, types, and footprints. For types, we need to interact with the ontology community.
- Wikipedia is a good way to solicit knowledge, but the method should not be used as a means for building gazetteers; rather, as a means for data collection. An organized research group is needed to put the harvested results into a structured and consistent format.
- Geocaching as model: why don't we have similar types of games for placenames to get people more involved in a more active way in raw data collection?
- We need a knowledge model that is adaptable to different forms of placename knowledge.
- Controversial thoughts for the sake of discussion:
    - Accuracy is more important than precision for geospatially referencing places. It is better to direct the user to a vague, general location than to a very precise footprint or to the wrong location entirely. It may be enough to know only that a place is within a broader feature.
    - Relative location is more important than absolute location. Most users do not care about the exact coordinates but they are interested in whether the location is near a river or atop a mountain.
    - Implicit local meaning and stories for places are more important than the explicit descriptors. Which country a place is in is less important than the location of major cities along a highway.
    - Place information provided by a gazetteer needs to be scale dependent.
    - When is folk knowledge considered acceptable for addition to gazetteers?

Sussion 2 Summary

David Bodenhamer (IUPUI)

- Things are always more complex than they seem.
- The Web is a sandbox, but what are the methods for using that sand?
- If building for everyone, are we building for no-one? What's our purpose in creating gazetteers?
- Do we need a scale-dependent knowledge model?

Session 3: Interoperable Gazetteer Services

Session chair: James Reid, EDINA, University of Edinburgh

Greg Janée

University of California, Santa Barbara: *Rethinking gazetteers and interoperability*
ADL Gazetteer Protocol

The motivation for the development to the ADL Gazetteer Protocol was to define the gazetteer's role in the ADL architecture. It was created as a co-development with ESRI in 2001; minor changes were made in 2003. It defines a simplified gazetteer model with a focus on interoperability; it is compatible with and mappable from the ADL Gazetteer Content Standard (GCS). It is a "lite" schema containing only the GCS required elements. The protocol client is Web-based. It supports seven query types. It is an abstract specification with an HTTP+XML instantiation. There is also a separate thesaurus protocol based on the Z39.19 thesaurus model to support the interoperability of typing schemes.

A gazetteer is defined as a set of gazetteer entries. Each entry contains an identifier, 0+ place codes (e.g., FIPS codes), a temporal status for the place (former, current, or proposed), 1+ names, 1+ footprints, and 1+ types. Type terms are linked to type schemes (e.g., thesauri). Names, footprints and types are all given a temporal status of former, current, or proposed and one value for each is designated as "primary". Gazetteer entries can be related to one anther with relationships such as "part of"; e.g., Santa Barbara is part of California.

There is no support for qualified placename queries; e.g., "find Santa Barbara, CA". It is onerous to implement low-level facilities to execute such a query, but it can be done with compound queries that search for "Santa Barbara" spatially contained within the footprint of "California" or "Santa Barbara" that has an explicit *part of* relationship with "California". The results of such queries are implementation-specific, unpredictable, and variable.

Interoperability Use Cases

Harvesting: This is the process of aggregating data from distributed gazetteers (especially local gazetteers). The process needs to be supported by protocols and representation standard(s). The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is an existing, operational harvesting protocol that could be used for this purpose; the GCS is a representation standard for gazetteers.
Lookup: Finding a place by name or other descriptive data. Examples shown of search parameters broader than a gazetteer:

- Google Maps: "go to a location," "find a business," "get directions"
- Emergency Contraception: Search for a provider: search by zip code or area code
- Global Land Cover Facility, Earth Science Data Interface: search by sensor, "WRS," start and end date, new since date, path, row, ….
- A site that supported lookup by street address, city name, zip code, latitude/longitude, and the URL of a flickr photo that has geotags.
- A site that shows gas prices by zip code.

Reverse lookup: Finding nearby places, nearby places of a given type, and nearest place of a given type.
Geoparsing: identifying and geolocating place references in documents. For example, the GeoNames.org RSS-to-GeoRSS converter. Geoparsing uses gazetteers, but not a gazetteer protocol.

Ontology-based reasoning: inferencing over a knowledge base of places; a unification of facts. Requirements include unique IDs for places and an ontology of relationships.

Gazetteer Protocol, Revisited

How well does the ADL Gazetteer Protocol meet these use cases?

- Harvesting: already better supported by OAI-PMH
- Lookup: too limited, rigid
- Reverse lookup: supports *near*, but not *nearest*
- Geoparsing: N/A
- Ontology-based reasoning: N/A

Question for the Workshop

Should we rethink gazetteer interoperability?

- From: have an entity (gazetteer), have a protocol for accessing that entity
- To: multiple protocols oriented around use cases and functionality that various kinds of entities (gazetteers, geocoders, etc.) participate in and implement to varying degrees.

Comments and questions for Greg Janée

- Jim Frew: Well-illustrated distinction between gazetteers and geoservices. We can agree more readily on the data model than on common services and can argue that there is no common service.
- John Frank: Comment on the goal of having a "simple" protocol: what's the lowest common denominator – or minimal stack – that could support all of these specific needs and more? What's the minimal set of data to which more specific services could add?
- Jordan Hastings: Referring back to the semantic triangle for gazetteer data, can we think in terms of specifying any two of the corners and get the third?
- RE semantic triangle: think of the transformation at the center of the triangle for getting the whole triangle on the basis of a subset of its vertices.
- Services are "finders" and are not part of the gazetteer model. There is a ubiquitous need for finder services and gazetteers can support these services.
- James Reed: Mentioned web 2.0, folksonomy, and mashup. Also, REST-based geodelivery services and XML, SOAP, and UDDI, registry-based discovery. A lot of the slots are already available and could be applied to decomposed functional requirements based on use cases. What about the e-Science/Grid computing communities – web services within workflow engines that consume geographic information? Focusing on use cases will show us what people actually want. Part of the research agenda could focus on harmonizing existing (competing?) protocols and approaches.
- There are services that don't map well into the traditional gazetteer data model. A more restrictive definition of the term might leave more room for a description of the range of services.

- Linda Hill: We need to think about how to represent gazetteers as knowledge organization systems (KOS) or objects in a wider information environment. What metadata structures do we need for gazetteers? Also, keep in mind that the gazetteer model can be and is used for more than placename-oriented description. For example, it has been applied to named time periods and discussed as a model for data such as named geo-temporal events (e.g., hurricanes). We need a typology of gazetteers. We also need to keep in mind that many of the end users and local data sources we need to interact with are not within sophisticated computational environments. Another need we face is the development of a toolkit so that people can build, maintain, and serve out gazetteers.
- Ray Larson: Mentioned SRW (*Search/Retrieve Web service*) or SRU (*Search/Retrieve via URL*) for query and search activities and also the next generation of the Z39.50 protocol. Instead of inventing something new, piggy-back on what already exists.
- Dealing with qualified placenames: include in the gazetteer? Can be constructed and pre-indexed from the hierarchical relationships in the gazetteer.
- John Frank: Mentioned the attribution string in Metacarta APIs and the licensing thereof. Is there a calculus of creating a name for a location? If well-defined, people could publish functions that generate name strings. A gazetteer protocol that has a few elements in the data model as possible would be good.
- The problem of super models: are there bridges from one to the other?
- The minimal gazetteer data model is use case dependent.
- Other ways to use the gazetteer model – a model for named things; an association between labels for things and coordinates. It can be used for events. Are addresses a form of naming?
- John Frank: Pointed to the geo microformats page: http://microformats.org/wiki/geo and suggested that a microformat could be made for gazetteer data.

Ruth Mostern

University of California, Merced: *A Historian's Perspective on Georeferencing and Interoperability*

Digital gazetteers are structured dictionaries of named places. They can be integrated with network-accessible service to respond to queries and retrieve geometries. They are also valuable information resources in their own right and they provide the basis for place-based search, display, and integration.

Another perspective on place can be found in *Place: An Experiential Perspective*, by Yi-Fu Tuan (1975).

- "Geographers have approached the study of place from two main perspectives; place as a location, a unit within a hierarchy of units in space; and place as a unique artifact…. [In the second sense,] place is a center of meaning constructed by experience."

- "Experience constructs places at different scales. The fireplace and the home are both places. Neighborhood, town, and city are places; a distinctive region is a place, and so is a nation…. They are all centers of meaning to individuals and to groups. As centers of meaning, the number of places in the world is enormous, and cannot be contained in the largest gazetteer."

In the *Foguang Encyclopedia* (Chinese sacred geography), the entry for Emei Mountain contains 25 named places in addition to "Emei Mountain" and an additional 210 locations on the mountain are referenced by category. The encyclopedia defines the mountain in political space and includes four alternative names (including Buddhist and Daoist names). It notes neighboring topographic features and names five of its peaks. It is part of a group of Four Great Sacred Mountains; all of them are named. Significant built structures on the mountain are named: five temples, one statue and one road. More than 70 additional temples, more than 40 grottoes and caves, and more than 100 stone niches for carvings are referenced. The entry for one of its famous temples notes that it is part of "The Six Great Ancient Monasteries of Mount Emei" and gives three historical names; also notes seven named buildings, a named statue, and a named relic.

The *Yellow Beryl* is a chronicle of the Tibetan Geluk monastic state written in the late seventeenth century; it is organized by region. Descriptive information about the monastic orders includes hierarchies, landholdings, and activities; the description of activities provides feature typing for the monastic entities. Associated information includes the names of affiliated individuals, ritual practices, and events.

Computer applications in the humanities raise interesting challenges: "The twin computational requirements of complete explicitness and absolute consistency open up a space for the scholar to refine an inevitable mismatch between a representation and reality" (Willard McCarty, 2005). Implications for gazetteer design include:

- supporting multilingual data
- dealing with intricate temporal dimensions
- handling spatial and temporal ambiguity
- using richly attributed textual sources
- reflecting indigenous ideas of scale
- including rich sets of relationships

The gazetteer model can be applied to named time periods and events also: *time-ateers*. Events are described by both place and time dimensions which are tightly bound – place is modified by time and time is modified by place. TimeMap™ is a "time-based interactive mapping" system, developed by the Archeological Computing Laboratory at the University of Sydney, which provides a structure for and access to distributed spatio-temporal humanities datasets distributed around the world. Datasets with spatial coordinates and time ranges associated with the data can be discovered and mapped over the Internet with this software. This includes "events," such as birth and death dates and the places these occurred.

The Electronic Cultural Atlas Initiative (ECAI) has pioneered in supporting the development of single socially authored gazetteers. These are created by multiple validated contributors, with or without an overall editor and include an associated discussion site. Wiki technology is used, either as a free-for-all or under the control of an editor. Inevitably, most historical gazetteers will emerge as small, handmade, specialist artifacts. To put them together requires solving significant issues: incompatible typing schemes, unmanageable ambiguity, conflicting authorities, and achieving buy-in from the participants/creators.

Comments and questions for Ruth Mostern

- Randy Flynn: NGA has many other databases that we do not think of as gazetteers, but in some ways they really are and they can be the sources of information for relating places and events within a structured knowledge organization system using the gazetteer paradigm.
- In the context of emergency services, it is difficult to share the collective knowledge required without compromising security and privacy concerns. Buy in is also an issue here.
- In regard to the statement about "unmanageable ambiguity," ambiguity is complicated but it's not necessarily unmanageable. You need to research the ambiguity to figure out if you can manage it. Supposedly intractable situations may turn out not to be.
- Bruce Gittings: Place ambiguity is real. You may not be able to determine whether a place reference is to one or another mountain. These sorts of ambiguity are not fully tractable.
- Re: "time-ateers": this may be a way to get to a concept of the "minimal" gazetteer. And, in regard to the gazetteer being a "knowledge organization system," is the gazetteer one type of a more generic idea of modeling and structuring information in some domain? Is this a genre?
- Gazetteers are culturally grounded. Given that you can lie with maps, can you also lie with gazetteers? If so, what are the consequences for modeling? How do you put an empirical measure on cultural distortion? In a historical context, the practice if to maintain fidelity to original sources and make use of rich attribution to sources – you must indicate where you got the information and why you interpreted it the way you did. A variation on the trust issue. Metadata is a touchstone issue here. Do we think that gazetteers currently have adequate metadata; in many cases, they are arguably not rich enough in the regard.
- Historical cultural gazetteers are more qualitative and relationship-centric than other gazetteers. An ontological approach is attractive in this domain. Reasoning systems based on ontologies have been generally confined to "practical" uses with the AI/computer science field. Is there any knowledge organization support technology that's actually been applied to gazetteers?
    - The National Counterterrorism Center has an ontology that doesn't know enough about linkages. In their system, rich attribution for new pieces of information is required and is critical. There is also a police application system that is about who did what to whom, when and where, for the

purpose of solving crimes. Covers events, places, and relationships. E.g., Cop Link.

- Ruth Mostern: We need to think rigorously about relationships. We know how to cite the sources of objective information, but how do we attribute relationships so that we make them rich, interesting, consistent and useful.
- There are a number of national mapping agencies that are developing ontologies, using tools coming out the AI community (e.g., OWL). They are not deliberately addressing gazetteers, but they are building up the ontologies of types and relationships for the databases to which the gazetteers refer. There's a lot of potential synchrony here.
- Naicong Li: We are exploring these issues, with a focus on the relationships between objects in out tortoise habitat management project. Data includes images and text documents, as well as GIS data. We have been using a gazetteer modeled about the ADL Gazetteer Content Standard. Our user community needs more than the gazetteer provides. What is the relationship between the ontology and the gazetteer? In theory, the ontology can automatically generate and maintain the gazetteer out of the more comprehensive ontology.
- Are gazetteers infrastructure? Should we be thinking in terms of a functional decomposition of existing protocols to support a richer suite of gazetteer services? Is it a matter of putting an individual gazetteer into an atlas system – all custom-built stuff? This is a different question from Greg's use scenarios. How do we think locally and build globally?

Session 3 Responses

Paul Ell (Queen's University, Belfast): [Unfortunately, Tom Elliott did not take notes on Paul's response, being otherwise occupied by projection issues

Tom Elliott (University of North Carolina)
The Pleiades Project

Organized by the Ancient World Mapping Center at the University of North Carolina at Chapel Hill, U.S.A., Pleiades brings together a global community of scholars, students and enthusiasts to expand and enhance continually the information originally brought together by the Classical Atlas Project (1988-2000) to support the publication of the Barrington Atlas of the Greek and Roman World (R.J.A. Talbert, ed., Princeton, 2000). Our name, "Pleiades" (the daughters of Atlas in Greek Mythology) reflects both this heritage and the forward-looking goal of collaborative diversification.

Combining community approaches (like those used by Wikipedia) with academic-style editorial review, Pleiades enables anyone — from university professors to casual students of antiquity — to suggest updates to geographic names, descriptive essays, bibliographic references and geographic coordinates. Once vetted for accuracy and pertinence, these suggestions become a permanent, author-attributed part of future publications and data services.

Gazetteer interoperability axes

- systems ↔ systems
- datasets ↔ systems
- users ↔ information
- scale and size
- time
- modes of production

What are "interoperable gazetteers"?

- systems/services: assumes sustained hosting, funding, and management; implies agreement on data formats and conventions; both purpose and audience are factors that affect the potential of interoperability.
- datasets/documents: how do we achieve persistence of results? agreement on formats, conventions, purpose, and audience are also involved here.

The advantages of the publication model for long-term availability:

- permanent record of reference for scholarly work
- copies are identical in content
- the published form is fixed and outside the further control of the author, although authors can change views or correct errors through a further published work
- editions are unambiguously cited
- published works can be irrevocably and identically replicated

Formats and standards

Ideally, standards are kept as simple as possible, with no special formats, purpose-built tools or essential behaviors and they include computationally actionable citations/references. If possible, use existing standards: e.g., Dublin Core for metadata, RFC 4646 for identifying languages, TEI for bibliographic citations, IMS vdex for thesauri, GML or georss for coordinates, Atom XML for wrappers. Standards should say explicitly what they mean rather than assuming that users can figure things out, and they should be designed so that they degrade gracefully.

Harvesting concerns

Is it adequate to assume that harvesting is a services+protocols-based interaction between two systems?

Requirements for gazetteer tools

- Guidance and stand-alone tools for the creation of gazetteers
- Guidance and support for the publication of gazetteers
- Mechanisms for uptake and propagation of published gazetteers

How do you know when your gazetteer is done?

[See http://icon.stoa.org/trac/pleiades/wiki/DGRPSession3Response for URLs for existing relevant standards.]

Breakout Session I

Questions

1. What organizations, individuals, opinions are not represented?
    a. How can they be included?
2. Are the three topics (components, process, interoperability) appropriate?
    a. Are there other ways of organizing the field?
3. What are the fundamental questions, principles, concepts of the field?
    a. Is there a rigorous theory of gazetteers?
4. What gaps exist in our current knowledge?
    a. What research is needed?
    b. What developments are needed?

Group reports

Group 1, Board Room

Members: Anderson, Connelly, Driver, Fisher, Freeston, Guptill, Henk, Hobbs, Jones, Mani, Mostern, Stone, Tobler, Yao

1. Who
    a. Implementers
    b. Companies
    c. Geographers of place
    d. Ontologists
    e. Scientists using type sites
    f. Not just North America
2. Topics
    a. Tools/methods
    b. Content
    c. Purpose/use
    d. Theory
    e. Interoperability in ingrained into all of the topics
3. Fundamentals
    a. Is there a rigorous theory of gazetteers? NO – but YES for ADL
    b. What would a rigorous theory be?
    c. Temporality
    d. Modeling footprints
    e. What is a place?
    f. Populating and maintaining gazetteers
    g. Validating gazetteer data
    h. Interoperating
4. Gaps
    a. Minimum set of components
    b. Gazetteers as knowledge organization systems
    c. Agreement about when a gazetteer is not a gazetteer
    d. Role of qualitative information; i.e., georeferencing

e. Generic vs. purpose specific approaches

Group 2, Hunt Room

Members: Bodenhamer, Couclelis, Ell, Flynn, Frew, Hancock, G. Hill, Janée, Larson, Mark, Reid, Svensson, Veisze

1. Who
    a. Users
        i. Consumers: human and systems
        ii. Of Google, Yahoo, etc.
        iii. Emergency responders
    b. Standards representatives, e.g. ISO, OGC (some dissent)
    c. "Grid" community
    d. Communities
        i. Research (NSF)
        ii. Computer science
        iii. Requirement definers
    e. DOD intelligence
2. Topics
    a. Yes, the topics chosen for this workshop were appropriate.
    b. Applications of gazetteers; helps inform decisions on representation
3. Fundamentals
    a. Regulatory requirements vis-à-vis gazetteer standards
    b. Wide spectrum of gazetteer typology
    c. Least common denominator among candidates to be called a gazetteer
        i. Is a collection of street addresses a gazetteer?
        ii. "Purposeful collection"
        iii. Is the definition of a gazetteer application specific?
        iv. Is being "authoritative" a necessary or sufficient condition for gazetteers
    d. MyGazetteer the equivalent of MyDocuments
    e. Gazetteer function: disambiguation
    f. Comparing definitions of gazetteers: OED (geographic index or dictionary), ISO (adds "type", "class of feature"), United Nations (adds "variants of toponyms")
4. Gaps
    a. Working definitions of gazetteers
    b. Expansion of use cases
    c. APIs
    d. Test environments
    e. Funding sources
    f. Overall objectives

Group 3, Carriage Room

Members: Carver, Dobson, Elliott, Frank Gittings, Hastings, Janowicz, Li, Morin, Rewerts, Tefft Veregin, Yuan

1. Who
    a. Not represented (whether aware or not):
        i. Google, Yahoo, MS, travel companies (e.g. Expedia), cognitive scientists (e.g., psychologists and linguists)
        ii. Web tools: collaborative atlases, wiki applications, social nets, geo 'communities'
        iii. Census
    b. How to: coding projects/open source
    c. Bigger hall? Alternative meeting structures
2. Topics
    a. Components of gazetteers; define the core
    b. Process (of georeferencing)
    c. Interoperability
    d. Abstract
        i. "Lightening talks" – many short presentations
        ii. Collaboration: peer-to-peer, wiki, …
        iii. Theory of locative structure/tech
        iv. Knowledge representation
        v. Authority; provenance
3. Fundamentals
    a. No rigorous theory of gazetteers – as opposed to definition
    b. Theory for what? common ground, harvesting, efficiency
    c. More than placenames, but placenames are the base element
    d. Sharing data – motivated by ??
    e. Amalgamation potential
    f. Guiding principles/ best practices
4. Gaps
    a. Use cases
        i. List of applications
        ii. Ecology of gazetteers would help
        iii. "Zoo"
    b. Where are the limits?
    c. A definition (laughter)
    d. Tools beyond parsing
        i. Comparative, analytical
        ii. Extract from GIS data
    e. Reaching users (contributors)
    f. Society
    g. Trust

Breakout Session 2

Questions

1. What are the priority topics for long-term basic research?
2. What are priority topics for short-term research?
3. What do we need to say to the funding agencies?
4. What is the future of gazetteer standards, protocols?
5. What activities would continue and strengthen the momentum of this meeting, help to build an effective research community?

Group reports

Group 1, Board Room

Members: Anderson, Bodenhamer, Carver, Fisher, Flynn, Frank, Henk, G. Hill, Mark, Morin, Mostern, Yao

1. Long-term basic research
    a. Ontology of footprints (representing and reasoning)
    b. Psychological research on human georeferencing categorization (cross-cultural, cross-lingual, political)
2. Short-term research
    a. Enumerating/characterizing "gazetteer" requirements
    b. Interoperability among very different gazetteers and data sets
    c. Integration strategies: fusion, conflation, X-walks
    d. Multilinguality (Romanization, transliteration, transcription)
    e. Taxonomy of users and requirements (use cases)
    f. Basic elements/minimum set for gazetteers
    g. Taxonomy of gazetteers
    h. Community repositories and wikis
3. Funding agencies
    a. Research will lead to useful applications (e.g., geographic information retrieval and location-based services)
    b. Where do gazetteers and other technologies work together?
4. Gazetteer standards and protocols
    a. Future TBD, based on use-case analyses
    b. Trust and provenance (structure used tradecraft)
5. Continuing activities
    a. Community tool repository and wiki
    b. Charettes: focused session involving designers and users
    c. Conferences: COSIT?; UN Conference on Standardization of Geographic Names

Group 2, Hunt Room

Members: Connelly, Couclelis, Dobson, Freeston, Frew, Gittings, Hobbs, Janée, Janowicz, Reid, Rewerts, Stone, Veisze, Yuan

1. Long-term basic research
    a. Role of the gazetteer as a knowledge organization system
    b. Vision, evolutionary direction
    c. "Global description service"
    d. Methodology for data acquisition
    e. Concept strata
    f. Scale-dependence; DNS-like model
    g. Liability issues; IPR
        i. wiki context
2. Short-term research
    a. Interoperability, test environments for
    b. Industrial benefit
    c. Use cases, collaborative development of
    d. Definition of micro-formats
    e. Implementation of existing OGC gazetteer specifications
    f. Sustainability of gazetteers patents and related GI
3. Funding agencies
    a. "Organize your knowledge"
    b. "Get the right people working together"; interdisciplinary
    c. "Add the 'where' to 'what'"
    d. Build an infrastructure for long-term, "big" science
    e. Avoid "digital earth winter"
    f. Workshops for defining new agenda for gazetteers
    g. Argue gazetteers as enabling technology
    h. Classes of fundable research: (1) important; (2) tractable; (3) not trivial
4. Gazetteer standards and protocols
    a. Local protocols; free forms for local knowledge leading to web-specified interfaces
    b. Analysis of existing and emerging standards that will lead to articulation of points among them
    c. ISO blueprints for gazetteer expansion of scope
5. Continuing activities
    a. Forum for discussion and exchange
    b. Interoperability experiments involving alternative local designs
    c. A web presence: e.g., opengaz.org or .net
    d. Gazetteer tracks in geospatial conferences
    e. Build professional credential process

Group 3, Carriage Room

Members: Driver, Ell, Elliott, Frank, Guptill, Hancock, Jones, Larson, Li, Svensson, Tefft, Tobler, Veregin

1. Long-term basic research
    a. Spatial natural language is vague; how to automate understanding of it, including the role of context
        i. Engage people in natural language processing area
        ii. Looking beyond the word level to ambiguous meanings
        iii. Tools and algorithms for parsing and extracting semantic relationships
        iv. Bounding and processing spatial phrases - moving beyond entity identifications
    b. The nature of footprints
        i. Fundamental geometric issues (accuracy) - what is appropriate to reference, and how
        ii. How defined
        iii. How deal with topologically referenced footprints
    c. Support for multilinguality
        i. User languages - multilingual gazetteers - multilingual access to geographic data or data leveraged via language
        ii. This is a problem of 2 parts
            1. Alternate names and spellings
            2. Scripts and transliterations
    d. Time
    e. Data acquisition and integration
    f. How to enrich the gazetteer, esp. the type thesaurus (the knowledge organization structure) so we can capture and related to other types of information, including relationships and object-intrinsic properties (e.g., county objects and attributes for recording their area)
2. Short-term research
    a. Getting real gazetteers for real people to do real things with
    b. Aggregating search results from multiple gazetteers - methods
    c. **Real** specific use cases, tied to test cases
    d. What is purpose specific and what is generic - most discussion today has been purpose-specific
        i. How do we get to the generic? - better answered from experience
    e. Inventory of existing gazetteers in computer accessible form
        i. Purpose
        ii. Content
        iii. Is there limited overlap?
        iv. We don't understand what's there
    f. Conference idea: real-life examples of gazetteer use
        i. Here's what we're doing
        ii. Problems /solution
        iii. Here are the gazetteers we're hitting (real data)

        iv. Objective: are there common tasks and methods - what are they - core functional requirements of gazetteer services (?)
    g. Creation of gazetteers
        i. Cheaper ways to populate gazetteers
        ii. Obsolesence
    h. Tools
    i. Fundamentals, both on data and interface sides, are not jelled
        i. Need to find common ground where both public and private sector interests intersect
        ii. Lots of beneficiaries
    j. Community-based, social models for data acquisition and refinement
    k. Low-hanging fruit: produce gazetteer entries easily out of existing GIS data- as long as those records have a name
        i. What are the principles for storing data in multiple ways - what are the boundaries there?
3. Funding agencies
    a. Government; research; commercial
    b. Incentives more consistent for getting behaviors than policies
    c. Assumption is this will be paid for by government?
    d. What are ways to make these things worthwhile to get them paid for in other ways than government funding
    e. Motivation in terms of major beneficiaries
    f. Major emphasis: what is the **point** of doing this - what are likely benefits
    g. There is a public requirement for efficient access to geographically referenced information
        i. This spans commercial, academic and governmental interests
    h. Long-term engagement is needed
    i. Deliverables to private sector need to engage and fuel them for further dialog
4. Gazetteer standards and protocols
    a. Why we need those standards? geographic information retrieval systems on the web capable of recognizing these names
        i. How will this information be maintained
    b. Rephrase question: what will drive the adoption of gazetteer standards and protocols?
        i. There's a strong motivation for something common to emerge
            1. Interoperability
            2. An all-knowing future thing
        ii. Emergence of a dominating gazetteer database that everybody wants to use - critical mass is essential
            1. But how do you keep it current
        iii. A model: used book sellers who advertise their books on about 3 websites
        iv. Another: sabre and the other one
        v. Microprocessor world for getting specs out

5. Continuing activities
    a. Funding
    b. Charges for working groups
        i. Surely there are small consortia within this community that would be interested in doing useful things
    c. Figure out who cares about these activities and why do they care?
        i. Business/mission/use cases will flow from these
        ii. Grounding this work in real examples is important
    d. An effective research community (a bit more inclusive) requires deliverables that are clear, marketable and palatable to those organizations/communities that we want to be part of this
        i. A SIG?
        ii. More frequent meetings?
    e. Focus on finding a common purpose: a gazetteer that we could all benefit from
        i. Is the common ground closer to the data or the techniques and approaches
            1. Try to find a common data set - a killer gazetteer
    f. Google Earth (and the like) is perceived by a large and growing number of people (decision makers) as **the** earth reference system for information retrieval
        i. Involvement of Google and the like, or use of the interface, would help
        ii. Identify their agenda and peer with them

Research Agenda

After the workshop, the following research agenda was developed based on the presentations and discussions during the workshop—in particular, on the ideas reported from the second Breakout session in response to the question about priority topics for long-term and short-term basic research. A draft version of the agenda was circulated to participants and comments were incorporated into this final version.

1. **Spatial and temporal modeling of gazetteer data**
    a. Modeling the inherent uncertainty of gazetteer data – toponymic, spatial, temporal, and classification (type) data, in storage, retrieval, and presentation
        i. Investigating the suitability of traditional fuzzy techniques for gazetteer data
        ii. Building statistical models of uncertain gazetteer data in order to reason about spatial references
    b. Modeling the temporality of places and their footprints, categories, names, and relationships
        i. Modeling of the temporal and spatial components of gazetteer data for applications in which the temporal aspects are on equal footing with the spatial (e.g., weather events, historical patterns, both

social and environmental)

2. **Ontological aspects of gazetteers and gazetteer services**
    a. Analyzing the options for cross-walking among feature/place categorization schemes and ontologies designed for gazetteer data
        i. Correlating place category schemes among languages, for a fixed set of features
    b. Developing and testing automatic methods for deducing place classification from placenames to create a 'gazetteer classification advisory service' that can be used to support gazetteer search interactions across distributed gazetteer services and gazetteer creation
    c. Developing automated methods for inferring placenames and types from topographic evidence (e.g., unnamed creek running beside Cobbs Creek Road)
    d. Developing comprehensive ontologies for gazetteer data; specifically, documenting the expected properties of various categories of features, as well as expected relationships among the categories, with properties for those relationships where appropriate.

3. **Geographic information retrieva**l
    a. Creating an information retrieval testbed environment where footprint generalizations and similarity calculations can be tested for performance for given tasks, answering, for example, when bounding boxes and convex hulls are sufficient are for geospatial information retrieval
    b. Testing the efficiency and efficacy of various areal footprint encodings: bounding box, enclosing circles, convex hulls, detailed GIS geometries (e.g., shapefiles) for gazetteer services

4. **Phonic aspects of gazetteers and gazetteer services**
    a. Developing techniques for the phonic representation of placenames in databases and as guides to pronunciation, including concern for grammatical context
    b. Investigating techniques for identifying/matching placenames phonically (e.g. soundex) as well as textually
    c. Developing prototypes of user interfaces to gazetteer services that use phonic modalities for both local and foreign users

5. **Placename intelligence** (or **Georeferencing intelligence**), especially in unstructured sources of gazetteer data and georeferences
    a. Recognizing references to spatial and temporal relationships in free text, parsing them, and translating them to geospatial coordinates and time ranges for spatial and temporal visualization and analysis
        i. Applying parsing techniques for spatial and temporal references to voice/audio data
    b. Capturing and manipulating topological references to place without metric support (i.e., without coordinate location)

c. Conducting a state-of-the-art study of geoparsing techniques and publishing the results
d. Developing on-demand procedures or hyperlinks to retrieve information from outside the gazetteer to complement gazetteer contents

6. Conflation of gazetteers and gazetteer data

a. Creating and testing techniques for the conflation of gazetteer data from multiple sources for one place with desired levels of confidence
    i. Developing new techniques for duplicate detection, particularly when attribution is inconsistent/sparse/indefinite (with or without context information), including the use of multiple sources of data to determine attributes of variant representations and associated uses/applications
b. Using placenames as a case study for combining data of mixed accuracies and precision, mixed granularities (e.g., names are coarsely-grained and vectors fine-grained), and mixed provenances
c. Developing methods for estimating the accuracy of gazetteer entries, before and after conflation

7. Gazetteer services and interoperability

a. Developing the specifications for software architectures and message structures for distributed gazetteer services, including distributed entry/editing of gazetteer data and harvesting gazetteer data for local purposes
b. Establishing a network of distributed gazetteers as a testing environment for gazetteer service interoperability, testing gazetteer service protocols and the suite of services needed to support discovery, search and retrieval
c. Harmonizing existing protocols and approaches to web-based georeferencing services and to gazetteer models

8. Gazetteer creation and maintenance

a. Developing a typology for gazetteers as knowledge organization systems and a metadata structure for documenting gazetteers
b. Automating gazetteer creation through the harvesting of geographic data from multiple non-gazetteer sources
c. Investigating methods for archiving and versioning gazetteers
d. Developing software for gazetteer creation and maintenance based on community standards that can be customized for individual and organizational purposes
e. Building specialized gazetteers from historical sources

9. Research into the naming process

a. Researching the types of features that are named in different cultures
b. Researching the processes by which names change during political and hegemonic transitions

c. Researching the roles of economic, religious, cultural, political, environmental, and linguistic factors in naming
d. Research into the processes by which names are maintained while their footprints change, evolve, or are (partially) undefined
e. Research into the trends in administrative/official naming procedures
f. Research into the commemorative and personal naming of places, buildings, airports, and the like

10. Users of gazetteers and gazetteer services

a. Conducting user studies for gazetteer services
b. Conducting a survey of use cases for gazetteer data

Participants

Dave Anderson	MITRE Corp
David Bodenhamer	IUPUI
Allen Carroll	National Geographic Society
Larry Carver	UC Santa Barbara
Thomas Connelly	Biblioteca del Congreso Nacional, Chile
Helen Couclelis	UC Santa Barbara
Mike Dobson	Telemapics
Beth Driver	NGA
Paul Ell	Queen's University
Tom Elliott	University of North Carolina, Chapel Hill
Peter Fisher	City University
Randy Flynn	NGA
John Frank	MetaCarta Inc
Mike Freeston	UC Santa Barbara
Jim Frew	UC Santa Barbara
Bruce Gittings	University of Edinburgh
Michael Goodchild	UC Santa Barbara
Steve Guptill	US Geological Survey
Lee Hancock	go2 Directory Systems
Jordan Hastings	UC Santa Barbara
Jordan Henk	Redlands Institute
Greg Hill	University of Colorado
Linda Hill	UC Santa Barbara
Jerry Hobbs	University of Southern California
Greg Janée	UC Santa Barbara
Krzysztof Janowicz	University of Muenster
Chris Jones	University of Cardiff
Ray Larson	UC Berkeley
Naicong Li	Redlands Institute
Inderjeet Mani	MITRE Corp
David Mark	University at Buffalo

Marc-Andre Morin	Defence R&D Canada
Ruth Mostern	UC Merced
James Reid	University of Edinburgh
Chris Rewerts	Army Research Office
Susan Stone	UC Berkeley
Bjorn Svensson	ESRI
Will Tefft	Maplink
Waldo Tobler	UC Santa Barbara
Paul Veisze	CA Office of Emergency Services
Howard Veregin	Rand McNally
Xiaobai Angela Yao	University of Georgia
May Yuan	University of Oklahoma

Acknowledgements

This workshop was funded by the National Geospatial-Intelligence Agency (NGA) and the Army Research Office (ARO) through an award to the University of Redlands (Mark Kumler, PI). It was convened by the National Center for Geographic Information & Analysis (NCGIA) at the University of California at Santa Barbara (UCSB) and the Redlands Institute in Redlands, California. The planning and execution of the workshop was led by Jordan Hastings, PhD candidate in the Geography Department of UCSB, assisted by Matt Rice, post-doc in the Geography Department of UCSB. Also assisting were the following Geography Department graduate students: Alan Glennon and Karl Grossner. Workshop participant Tom Elliott created and posted excellent notes at http://icon.stoa.org/trac/pleiades/wiki/DigitalGazetteerResearchAndPractice, which were depended on heavily for reporting on the session responses and discussions.