Digital Gazetteer Research & Practice Workshop

Position Papers

December 7-9, 2006 Upham Hotel, Santa Barbara, California

Introduction

In late 2006, the National Center for Geographic Information and Analysis (NCGIA) and the Redlands Institute held a two-and-a-half day workshop in Santa Barbara focused on the role of digital gazetteers in georeferencing applications. The workshop started with overviews of the state-of-the-art and current activities and lead to a consensus on the opportunities and directions for collaboration and the advancement of a research and practice agenda.

This document contains the list of workshop participants, the position papers submitted by these participants where applicable, and the meeting schedule. The final meeting report can be found in a separate document in this archive.

Participants

Dave Anderson

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
Telemapic

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

Gregory 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

Position Papers

(In Alphabetical Order)

STATEMENT OF INTEREST

TO: Organizing Committee
Digital Gazetteer Research and Practice Workshop

FROM: David J. Bodenhamer
Executive Director, The Polis Center at IUPUI, and Professor of History

RE: Workshop Application

DATE: September 26, 2006

Please accept this memo as a strong expression of interest in the December 7-9, 2006, workshop on digital gazetteers. The opportunity could not be more timely nor my need greater.

Over the past decade—and especially over the last five years—I have been involved in a variety of projects related to the digital humanities. Most of these activities have involved electronic cultural atlases and historical GIS, especially the development of Web-based projects (e.g., the North American Religion Atlas, www.religionatlas.org). This work, in turn, has led to what is currently a loosely federated effort with EU partners to create a conceptual framework for an initiative to link various national historical GIS efforts. It also has resulted in a parallel but related effort to use GIS to locate and link an array of UK digital historical resources in multiple data formats. This effort, which involves several colleagues from the UK, will rely heavily on the use of digital gazetteers, some extant and others that must be created. My compelling interest in this workshop is to ensure that we enter our work on both fronts fully aware of the advances made by groups such as the National Center for Geographic Information and Analysis.

As the director of a center that sues GIS as its technology of choice, I also have a strong desire to know about the potential value of digital gazetteer research for our larger agenda. The Polis Center at Indiana University Purdue University Indianapolis, established in 1989, has developed a strong record for applying GIS effectively at local, state, and federal levels. A completely self-funded unit with 25 FTE staff, we have developed a comprehensive community information system for the Indianapolis MSA (www.savi.org), an internal Web-based mapping and evaluation center for the National library of Medicine and its National Network of Libraries of Medicine, and, in a work-in-progress, a system of distributed and integrated Web services for mapping and spatial analysis for a variety of state agencies and counties in Indiana. We also work extensively with federal and state emergency management agencies on disaster planning, including significant involvement in training users nationally on HAZUS, a GIS software developed for pre-disaster mitigation efforts. Finally, over the past several years we have increasingly teamed with researchers in the IU School of Medicine, located on our campus, to incorporate spatial analysis in public health-related projects. In all of this work, it is becoming clear that digital gazetteers are important resources that we must learn to tap effectively.

The UK and EU projects represent the more innovative scholarly applications of spatial technologies and are the ones that most immediately require the creative use of digital gazetteers. In the UK effort, we are working with colleagues at The Queens University of Belfast, University of Lancaster, and University of Nottingham to use a digital prototype of the Survey of English Place Names, with x-y coordinates, to link of a variety of digital records (texts, databases, and images) available in the archives of the Arts and Humanities Data Service of Great Britain. Our initial aim is to use GIS as the means of searching, integrating, and visualizing these data. The challenges are complex, not the least of which is how to manage name changes across time and how to use these names and their variants to geoparse the other resources. We also wish to integrate the historical gazetteer with standard gazetteers, such as Getty or EDINA, and to define a solution to the lack of comprehensive time-period crosswalks. This workshop promises to help us avoid both conceptual and design errors.

The EU project is not as advanced as the other project, but it too will involve both historical and contemporary place-name gazetteers, although the issues will be even more complex because of language and classification differences. Numerous national historical GIS projects are in the works or completed, including Great Britain, Ireland, Germany, Netherlands, and Spain. Designing the spatio-temporal framework to handle continuously shifting boundaries is a daunting task in itself, and this matter is complicated by the problem of handling place-names in contested spaces, a phenomenon largely unknown in U.S. history. Again, attendance at this workshop would be useful in helping think through the options for addressing these conceptual and methodological issues.

Contact Information:

David J. Bodenhamer
The Polis Center at IUPUI

Position Paper
Digital Gazetteer Research and Practice workshop
November 17, 2006

A Case-Study-In-Progress: How a Media Organization Tackles the Georeferencing Challenge/Opportunity

Founded in 1888 to increase and diffuse geographic knowledge, the National Geographic Society has grown into a multi-faceted organization producing editorial material in multiple media and across numerous business units. The Society, like other media companies, is working hard to make its digital assets more versatile and accessible amidst a rapidly evolving, increasingly fragmented media marketplace. The growing popularity of Web-based mapping platforms and consumer-oriented GPS applications have spotlighted a particular opportunity, namely organizing NGS content for presentation in the context of location. The Society's venerable brand, its renowned cartography, much of its editorial content, and the core of its mission are all about geography. For all these reasons, the notion of georeferencing Society content has been an easy sell.

The process of *implementing* georeferencing, however, has proven to be anything but simple. Challenges include the vast scope and variety of content at the Society, its archiving in scattered locations across multiple business units, its access via a variety of databases with little or no common standards or nomenclature, and widely varying rights restrictions. Despite these challenges, progress is being made in the creation of an enterprise-wide infrastructure, and in pursuing project-based initiatives fueled by specific business opportunities, most of which complement and strengthen the enterprise effort.

Overall goals of the enterprise-wide effort are to create an infrastructure for organizing and accessing content by geography, including archival content and new content as it is created in the field or archived at the office. The infrastructure must allow for a variety of media formats, several database and metadata systems, and a range of potential applications including print, Internet, video, film, and mobile/GPS. An additional goal is to integrate georeferenced content with National Geographic's cartography, potentially providing a unique competitive advantage.

The key to georeferencing legacy content, of course, is the placename or names associated with that content. National Geographic Maps' placenames database, which is essentially the same as the index to the Eighth Edition Atlas of the World, was identified as the basis for a "master gazetteer" against which other databases would be "harmonized" or cross-walked. Disadvantages of the atlas index are the small number of placenames (some 140,000) compared with other gazetteers, and the fact that the features those placenames are associated with are cartographic (adjusted for visual display) rather than truly spatial. Advantages include the high editorial quality of the placenames and their association with a

GIS and its attendant schema (hierarchies, feature types, etc.).

Off-the-shelf software will be utilized to cross-walk or conflate NGMaps' placenames with at least three other NGS databases, each with more than 20,000 entries: the Society's publications index, the Film Library database (video and film), and the Image Collection database (photography). The Getty gazetteer will be tapped as a reference tool to identify placement within the partitive hierarchy for those placenames not in the current Maps database. Meanwhile, the hierarchical structure of the Maps placenames index (continent-country-province, etc.) will be enhanced using standards and practices of existing online gazetteer services, but fine-tuned to suit the special requirements of NGS. Placenames will be matched with lat-long coordinates (points, lines, bounding boxes, and/or polygons), at which point it will become a bona fide gazetteer. Finally, the same placenames will be linked to a Web-enabled version of the new seamless, multi-scale GIS cartographic databases currently under development in NGMaps, making it a *GIS gazetteer* as well. The master gazetteer will be accessible within NGS via a Web interface, enabling archivists to use it as a reference tool for georeferencing new and future content. Thus over time the need for cross-walking will decline.

Inevitably, the Society will embrace the use of GPS-enabled cameras and video equipment. Meanwhile, NGMaps is using software tools to incorporate spatial information into image headers and metadata, tapping placenames, addresses, and descriptive information. National Geographic Maps, the Society's Digital Media group, and the Library and Information Services division are in discussions with MetaCarta about using their powerful text parsing tools to facilitate the georeferencing of text-based content and media metadata.

NGMaps and Digital Media have been collaborating with ESRI to incorporate georeferenced content into the MapMachine, National Geographic's mapping platform. The MapMachine will soon be able to display icons that will provide access to hundreds of photographs, articles, video clips, and sound files. This will provide a new means of access to the vast assets of the National Geographic website. It will also serve as a proof of concept and, we think, a beta version of an exciting new storytelling tool. For 91 years, cartographers at National Geographic have used maps to tell stories about the world. Linking dynamic maps and multimedia content through georeferencing will enable us to create authoritative, curated resources, including narratives, guided tours, news features, and curriculum materials, bringing our spatial storytelling skills to bear in new media environments for the benefit of a global audience.

Allen Carroll
Chief Cartographer
National Geographic Society

Digital Gazetteer Research and Practice

Expression of interest

20 Sept. 2006

Thomas Connelly
Researcher
Sistema Integrado de Información Territorial
Biblioteca del Congreso Nacional
Chile

Chilean context

The benefits of georeferencing information have become increasingly evident over the last five or six years. Indeed, the team tasked with coordinating government agencies in this regard (SNIT: Sistema Nacional de Información Territorial) has recently been given a new, more specific action brief. This growing awareness of the importance of the adequate management of georeferenced data has created the opportunity to promote the adoption of the best criteria and practices of the international community. Thus, we face challenges on two fronts: how to learn those criteria and practices and how to best further their adaptation to Chilean reality.

The Library of the National Congress has pioneered the promotion of the design and implementation of national policy for the management of spatial information, focussing on gazetteers as key components in the creation of metadata records and offering a gazetteer-cum-thesaurus for use and comment via Internet.[1] In fact, the Library Web site offers the only place-name-based search tool available for Chile (apart from GoogleEarth, of course).[2] Although we do make maps for the specific purposes of Congress persons and our Web site allows user interaction with our data bases to create cartographic visualizations of statistical data, the Library is chiefly a consumer of georefrerenced data and, thus, a manager of territorial information.

We have come to realize that gazetteers are a powerful instrument to this end and participate in the Chilean inter-agency discussion on what their creation and maintainance imply and how to use them adequately.

Creating metadata records

[1] See: http://geoinfo.bcn.cl/ (in Spanish).
[2] See: http://www.bcn.cl/portada_siit.html (in Spanish).

Our interest in the Workshop revolves around the difficulties frequently encountered in the creation of metadata records that are only partially (or potentially) solved by the adequate creation of gazetteers, that is, how should the person creating the metadata record enter the Place Name and Theme Key words?

To best resolve this problem, we need to learn, understand and know how to implement the principles and farmeworks geared toward the appropriate management of georeferenced data.

Specifically, what is the relationship between the place name layer of data used to construct the map in question and the Place Name Keyword entries in the metadata record? Can the entire layer be inserted by machine into the appropriate fields of the metadata record? And, from the point of view of eventual queries (user searches), how can the feature type, as defined by the appropriate thesaurus, be related to the Theme Key word field? In other words, we feel that the creation of metadata records needs to be customizable and/or automated so that the ease with which they can be created will make them easy to handle with both geodata managers and consumers.

Within the context of gazetteer support for geographic information retreival systems, we are interested in the creation and sharing of category schemes for gazetteers, as related to the management of and access to georeferenced data objects.

With regard to georeferencing as a process, we wish to avoid beginners' mistakes and take advantage of accumulated knowledge to promote the adoption of current best practices for the naming and typing of places.

A central concern of ours is addressed in the area on interoperable gazetteer services: embedding gazetteer lookup capability in operational systems, esepcially insofar as this issue is related to problems of management and access to georeferenced information, specifically as regards the creation of metadata records and Internet user access to that information.

My experience with gazetteers has spanned my careers as an educator, executive of a mapping company and as a central component in my consulting practice (TeleMapics).

I have practical experience in creating geographic gazetteers, indexes and in the problems and potentials in deciding “appropriate’ name forms and publishing policies governing their use at Rand McNally & Company. Rand’s business was international and I was exposed to the pressures of “commercially-based name preferences”, as well as politically charged “naming” issues across the world. During my tenure at Rand, I licensed data to Getty for their Geographical Thesaurus and had discussions with a number of companies interested in geographic names and their use for purposes far beyond the world of maps. Perhaps of more interest, I was with the company when the name systems were automated for our internal use.

In my consulting career, which is focused on Local Search and Location Based Services, I have found that geographic gazetteers are an area of interest to all of the major players in Local Search, as well as those focused on general Internet search. In the last year, Yahoo acquired WhereOnEarth, a UK based company that created a unique digital gazetteer capable of aiding search engines to disambiguate geographical terms. The company’s products are now being integrated with Yahoo’s search engines in attempt both to sharpen the relevancy of results to queries that include geographic tokens and to assist advertising buyers with “buying” and targeting of advertisements that could enhance their sales on a geographic basis (and increase Yahoo’s profitability in the process).

I have spent considerable time, recently, examining many online gazetteers and speaking to developers of these products. In a make versus buy project for a client, I spent considerable time comparing gazetteers and creating a work plan for creating one from scratch using both proprietary and public domain sources.

Finally, over the last year, I have been involved as an expert witness is a patent case involving the use of gazetteers and potential intellectual property issues with certain geographic naming conventions.

The three focus areas of the meeting are especially relevant to my research interests (as much as a consultant can actually have “research interests”). I am interested in georeferencing as a process, especially in respect to its use on online search (including local search). Similarly, I am interested in issues surrounding geographic interoperability, as well as components of gazetteer services in respect to search.

I am intrigued by the potential opportunity to create a spatial version of UDDI to facilitate the creation of an international naming authority allowing “places” to list their details on the Internet. Not quite a WIKI perhaps, but something more effective than what we have today.

For these and many reasons, I would be interested in attending the Workshop in Gazetteers.

Mike Dobson

Statement of Interest

The Digital Gazetteer Research and Practice workshop interests me for two reasons: digital gazetteers afford the opportunity to expand both the nature and the extent of NGA's support to our customers; in addition, work on gazetteer questions affords an outstanding opportunity to deepen our collective understanding of the nature of and limits to formal descriptions of geospatial data and of commonality between expressions of geospatial knowledge and natural language. The questions posed in the workshop description illustrate the potential for addressing both "applied" and "fundamental" questions within a single, focused body of research.

Components of gazetteer services

The topics identified as focus areas for the workshop are likely to reveal a host of related questions, starting with "what is gazetteer data and, in a digital world, how does it differ from 'regular' geospatial data?" Does the difference lie in the data, in the assumptions that we make about it, or in the kinds of services offered? Which of the following would be gazetteer data: township names; zip codes; telephone dialing codes; designated service areas for commercial services; areas of responsibility for military or non-military organizations.

It might be useful to discuss the expanded utility of gazetteer data provided in digital form, along with scenarios for using new capabilities. Not only will digital services render it faster and easier to perform searches that theoretically might be done with hardcopy gazetteers, new capabilities might support entirely new services and expand the community of users for gazetteers. An extreme case might be to compare an excerpt of spoken language to a list of place names whose various pronunciations are indexed.

Such indexing might require capabilities well beyond today's soundex representations, e.g., application of phonological rules for multiple dialects, in addition to grammar and morphological rules. More straightforward uses might be to find spoken and orthographic variants of a place name (within a designated dialect or sub-dialect or across languages), to find alternative referents for the name and to find (or recognize) alternative names for the "same" place.

Georeferencing as a process

Many of the relationships between bits of linguistic data, and bits of geospatial data, are implicit and context-dependent; a better understanding of such relationships and how to exploit them is essential. Explicit identification of "all" relationships will fail for two reasons: first, the cost of capture and exploitation would be prohibitive; more importantly, we cannot enumerate all of the relationships of interest for a given bit of data.

Traditionally, we have looked for general rules and relationships in grammars and in sets of logical operators that can be combined in useful ways. Linguists impose an "adequacy requirement" on grammars that we do not use in computer science: a grammar should

generate all good constructions in the target language, and it should unfailingly identify faulty constructions as such. What is the adequacy criterion for our work?

Another line of inquiry might be methods and practices for populating and verifying gazetteer data. Are there common parameters we can use to set priorities for tapping into already-compiled (although perhaps not yet digitized) data or for organizing the work? Can we define methods for using the wide range of data becoming accessible to verify, validate, correct and organize place name information? Are there general practices that might be reliable, and can we combine atomic operations into tools to facilitate these practices? Are there ways to tap into user behaviors to find and resolve anomalies in the data? Can we define practical figures of merit for characterizing the quality of data in place name repositories? How do they use or differ from characterizations of quality in "traditional" geospatial data bases?

An ancillary question for anthropologists or geographers might be "how could we use these large, diverse repositories to enhance our understanding of how humans organize place?" In addition to identifying relationships that all cultures capture, are there "potential relationships" that are never used or that always require indirect reference? What is the significance of either finding?

Interoperable gazetteer services

Descriptions of universal underlying topologies for systems more straightforward than place names (e.g., kinship systems) have long eluded anthropologists and linguists. Thus, it is likely that gazetteers will have to be designed to support multiple organizing principles, including mechanisms for handling data that doesn't fit the schema in use. It might be useful to focus considerable attention on powerful, flexible tools for integrating data from multiple sources. It might also be useful to consider underlying principles for segregating and relating data that we might be able to use repeatedly.

Finally, it would be useful to organize the many classes of questions that this workshop will touch on to capture those where likely answers may lie in other disciplines, those that are dependent on one another, those that must be answered (or assumed), consciously or unconsciously, as gazetteers are built, maintained and used with today's tools, and research questions that are likely to enable great leaps of progress.

Beth H. Driver

Tom Elliott, Ph.D.
Director, Pleiades Project
Ancient World Mapping Center
University of North Carolina at Chapel Hill

I am interested in issues of collaborative geodata creation and maintenance, and interoperability of geoservices for humanistic research and teaching. I am particularly interested in the tension between new community-based approaches to content creation and well-founded traditional concerns about accuracy, citation and verifiability. Two other areas of concern are: cross-project collation and georeferencing of toponymic gazetteers, and feature extraction and georeferencing from historical maps and cartifacts that cannot be rubbersheeted usefully.

As the director of the Ancient World Mapping Center's Pleiades project, I preside over an effort to build a collaborative, web-based system for the perpetual update and diversification of a complex geo-historical dataset for Greek and Roman antiquity. The dataset itself derives from the work of the Classical Atlas Project (1988-2000). Comprising more than 50,000 named and unnamed geographic features, it was compiled and vetted by a 200-person international team of scholars and professional cartographers. Their initial goal was the creation of a scholarly reference atlas in print, and this goal was realized in 2000 with the publication of the *Barrington Atlas of the Greek and Roman World* (R. Talbert, ed., Princeton; see www.unc.edu/depts/cl_atlas). The Center's core mission, from its foundation in 2000, has been the perpetual update, diversification and dissemination of this legacy.

With initial funding from the National Endowment for the Humanities (2006-2008), Pleiades combines open, community-based content development approaches with rigorous editorial review. Pleiades will soon enable anyone — from scholars 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 will become a permanent, author-attributed part of future publications and data services. These will include OGC-compliant Web Mapping and Feature Services (WMS/WFS), as well as a geocoding/gazetteer service sensitive to the full range challenges posed by our toponyms: fragmentary witnesses, scholarly hesitation in assignment, variant orthographies and scripts, etc.

To support the Pleiades "community of practice" we are customizing an enterprise-quality content management system (plone.org) by adding custom "content types" to handle structured spatial, toponymic and bibliographic records (beta site with early results: icon.stoa.org/pleiades-beta). Robust version control, document history and threaded comment-and-review mechanisms will facilitate not only granular and incremental updating of individual records, but also large-scale expansion and diversification of our holdings. A flexible, dynamic mapping tool will permit on-the-fly visualization of arbitrary subsets of the dataset (including query results). Working groups will facilitate collaboration on topics of group interest.

Plans are now in preparation for a multi-year, collaborative effort to leverage this dataset and its maintenance environment to establish a reliable digital infrastructure for Greek and Roman geography. This effort will involve a number of major digital projects that are cataloging and publishing documentary and archaeological resources (e.g., for inscriptions, papyrus documents, coins), as well as critical reference resources (e.g., for personal names and prosopography). Most such projects in the field of Greek and Roman history have limited budgets and staff and therefore

cannot undertake to implement and maintain dynamic mapping and spatial query capabilities. Instead, in order to make best use of the planned Pleiades web services outlined above, other projects will need to upgrade their toponymic thesauri to include Pleiades ID references as foreign keys. Through an automated/supervised collation process, we will work with these projects to match their records against ours for each geographic feature of interest.

This is a formal request for a place at the Digital Gazetteer Research and Practice Workshop to be held at UC Santa Barbara. In addition to this expression of interest I attach a two-page résumé.

I am director of the Centre for Data Digitisation and Analysis at Queen's University Belfast. The Centre's main interest is the development of scholarly spatially referenced historical electronic research resources, and the exploitation of these large datasets in research using GISc. I have been involved in many large-scale projects relating to the construction of e-resources and their utilisation including the Great Britain Historical GIS, where I was a co-applicant on its main Economic and Social Research Council funding award; the associated 'popular' Vision of Britain Through Time incarnation of the GBH GIS with a variety of multimedia materials all referenced by place; the development of the Database of Irish Historical Statistics, which resulted in a significant publication mapping and analysing the geography of the 1840s Irish Famine; two new projects digitising, and associating by place name, medieval British sources including the 1086 Domesday Survey and a multitude of later medieval materials; a project which digitised and spatially referenced the first map of the whole of Britain – the Gough Map dating from around 1350; and the construction of three text-based resources containing geographical information – a virtual library of materials relating to the Act of Union between Britain and Ireland, a large repository of British and Irish Parliamentary Papers, and an e-version of the debates held in the Northern Ireland Stormont Parliament from 1921 to 1972. What is common to all these e-resources is that they can be referenced, and associated, by location.

In addition, I am co-author of a Cambridge University Press book that will be published in 2007 examining the use of GIS in history and geography research. I am an active member of the UC Berkeley based Electronic Cultural Atlas Initiative chairing ECAI's Scholarship and Content Committee and the E-Publications Committee. At Queen's University I head the Spatial Technologies Research Forum, a body created by Senate to promote GIS across the humanities and arts as well as the traditional disciplines in which the technology is used. I am editor of *Humanities Computing*, which supersedes the Edinburgh University Press journal *History and Computing*. The re-launched journal will have a strong multimedia GIS bias.

My primary interest in gazetteers is the potential to use them to associate disparate e-resources, such as those briefly described above, in space and time. As such elements of the workshop which focus the components of gazetteer services and on interoperability are particularly relevant. I recently received funding from the Arts and Humanities Research Council in the UK to hold a workshop on 'GIS e-Science' to discuss whether GIS technology might form the basis for organising and interrogating the rapidly developing range of e-resources in the Arts and Humanities, e-resources that are always, or almost always, referenced in space. It was quite clear from the workshop that if GIS was going to make a contribution to e-Science with relation to the Data Grid an effective way of linking information by location was vital. There are obviously many ways of expressing location as, for example, a defined polygon composed of administrative boundaries (as with the Great Britain Historical GIS or the Database of Irish Historical Statistics), a set of co-ordinates forming an arc or point (roads, rivers and castles as in my Gough Map project), or, and most commonly, an amorphous poorly conceived sense of location normally referred to by a place name (as with the Domesday GIS, the Medieval British Isles 'Domesday II' project, the various Parliamentary Paper projects etc).

In the UK I am leading a consortium (including AHDS, AHDS History, UC Berkeley, IUPUI, Portsmouth University and Nottingham University) that is making application to the Arts and Humanities Research Council under their e-Science grant call. AHRC have focussed the call on the development of e-Science infrastructure or substantial primary research using e-Science methodologies. The project I am working on does both. It will develop a comprehensive hierarchal place name gazetteer and use the gazetteer to advance

understanding of place-name geographies. The proposal will take the ongoing work of the English Place-Names Society (EPNS) which, since 1924, has been engaged in the painstaking collection and analysis of all the England's place-names, including the names of administrative units, settlement sites, topographical features, field-names and street-names. These names are extraordinarily revealing about the cultural and social patterns of English history: the suffix 'by', for instance, can be seen to chart the Scandinavian settlements of the ninth and tenth centuries, while Celtic survival in Anglo-Saxon England is marked by recurrent instances of Walton 'the settlement of the Welsh'. Topographical names and field-names record changes in landscape (e.g. extent of former woodland) and land-use. And so on. Fortunately for modern scholarship the EPNS has systematically collected place-name spellings from a wide array of textual sources: these spellings are arranged chronologically and related to their modern forms (where these survive). Shifting spellings of the same name are dated; wholesale replacements of the name are recorded. The history and derivation of the name is discussed and where possible explained. This material is printed in, to date, over 70 volumes of the Survey of English Place-Names, which are highly regarded as standard reference works. They are of immense value to the study of history, with either a national or local focus, language, geography, archaeology, family history and genealogy, literary studies, and many other disciplines that need to understand the geographical and environmental contexts of the subjects they study or the places they live and visit. Few reference works can claim such a wide audience.

One could not hope to find a more exhaustive or detailed gazetteer recording variant spellings as they appear in a vast array of historical documents, and arranged within a basic hierarchal structure. It is our intention to take the EPNS work and create an electronic gazetteer and use the rich holdings of AHDS to associate e-resources by location and chronology. Through a number of AHDS-held sources it will be possible to develop a range of potential footprints for the places recorded by EPNS – sources such as the Great Britain HGIS and Kain and Oliver's work which will provide polygons, vill co-ordinates developed from medieval Lay Subsidies and the *Taxatio* and point location files created by various Domesday scholars will allow the utilisation of various point coverages and Campbell's work on medieval England goes further and develops pseudo-polygons from point data.

Of less interest to the UCSB workshop, but to satisfy AHRC requirements, and demonstrate the importance of the gazetteer, the etymology and distribution of place names will be examined. Much work on this has been carried out by EPNS and others. What has not been possible in the past however, it the ability to link place names to a range of existing e-resources containing historical socio-economic and environmental data which can help to explain place name geographies. As well as demonstrating the potential of e-Science and the Data Grid, the project also provides an exemplar showing how gazetteers can associate sources and result in new scholarship.

Participation in the workshop will inform and enhance our AHRC grant application which will put gazetteers at the core of much research in the humanities and social sciences in the UK. It will allow me the opportunity to brief colleagues on our plans and discuss possible collaborations. It will also assist me to brief ECAI content affiliates of developments in the field. Finally, *Humanities Computing* would be extremely interested in featuring, at the very least, a report of the workshop in a future volume of the journal. To assist in my funding my attendance at the workshop, assistance towards the cost of travel and accommodation would be exceptionally helpful.

Where is that place? Modelling spatial footprints for a gazetteer

by Peter Fisher

Department of Information Science
City University, London EC1V 0HB

Historically gazetteers have been lists of place names which form the index of maps. Thus the classic gazetteer is bound as part of the atlas, sometimes as a single gazetteer for the whole Atlas and sometimes multiple gazetteers are bound in for different parts of the atlas whether single or multiple maps. The gazetteer is the principal means by which users of the atlas conduct guided searches to access the information in the atlas. Clearly another popular method for accessing the information is unguided search where the user has a general idea of the geographical region in which their place of interest is located and they use the map index sheets to find the correct map and then search on that map using local knowledge to assist refining the search. The only notion of space in these historical gazetteers is of a point or grid cell on a particular map; the point where the name of the place occurs or the cell which includes it.

A Gazeteer is becoming a list of the named features on a map and therefore is also an official or semi-official record of "place". But in society "places" are created and destroyed. I do not mean physically destroyed although that of course can happen but I mean rather erased from the cultural record – the name goes out of use. Indeed physical destruction does not mean that the name goes out of use.

But how are gazetteer entries collected? For example, I know that the Ordnance Survey of Great Britain originally collected names systematically from local worthies who assisted drawing sketch maps of the extent of the features. This information has never been digitised and is now lost to all but researchers. These sketch maps were used in deciding which names to include on maps and the hierarchy. This information is now being stored in the National Archive at Kew.

If the names on maps are an official recognition of places then why do names on maps come and go and come back through multiple editions of maps within a short period of time? Why do different editions of maps have different hierarchies of placenames encoded? Why do the positions of places in that hierarchy vary among maps? How is a particular hierarchy decided? Ultimately why is a particular name on any particular map?

Searching gazetteeers can produce ludicrous results for inquirers. If you look for the Marsall Islands in the gazetteer accompanying Google Earth, you do not see islands but an extent of water. If you search for Ilkley Moor in the Ordnance Survey electronic gazetteer, you are offered three different locations all within Ilkley Moor but in most interfaces you have to choose just one of them. Similarly if you search for a particular street in an A-Z atlas for the UK, it is referenced by a single cell on the map even if the street itself crosses multiple cells and even multiple maps. The same think happens on Multimap.com or Mapquest.com. The same problem was associated with the original GNIS where all rivers were located by a single location.

All these searches have one thing in common. A more flexible and appropriate response would not be to show the one index location for the feature in question, but a map that shows the full spatial extent of the feature. At the simplest the minimum bounding rectangle for the feature will provide this. To achieve this only requires that the every entry has two coordinate pairs instead of one.

I cannot be exhaustive, but I am aware that some modern gazetteers have gone further than this. The Getty Thesaurus of Geographic Names, for example, does not just have the name of inhabited places with a single grid reference but the hierarchy of named places in which they fall is stored.

Further enhancement of the gazetteer would be to have variable interpretations of what is meant by the term. So a query for "I want to see exactly the extent of the City of Leicester" would result in one response from a map server, while "Where is Leicester?" would result in a different map showing a more general context for the city. As long as the city is discernable on the map one that shows anything up to the full extent of the globe provides information to the inquirer. So should an

inquirer be specifying the extent of a context e.g. "where is Leicester in Europe?" What if the context is not known? What other information or locations should be shown on the map?

The question "I want to see exactly the extent of the City of Leicester" refers to a political entity which by fiat has definite boundaries, and so it is possible to define the extent of the political entity "The City of Leicester". However, does that mean that it corresponds to any individual's view of the extent of Leicester City or of Leicester? Did the questioner mean to refer to the political entity?

If we move to informally named locations the problem is more acute. Where, for example, is London's West End? Most people in Britain could give you an idea of where it is, and many Londoners (at least) would agree on a prototype for the West End, but many would not, and many would certainly not agree on were it extended to in any particular direction. How do we capture the extent of such poorly defined locations? Is Oxford Street in London more than the extent of the tarmac road which bears that name or does it include the shops which line it or even some number of the surrounding roads and the houses and shops along them?

Mountains and similar landform features are also poorly defined spatial entities about which I have written in the past (Fisher et al., 2004; Fisher et al., in press). Where is Helvellyn or Ben Nevis or Everest? The summit of each is clearly one definition of where it is, but the catchment area for it is another (upper and lower bounds?), but within that broad extent places nearer the summit are more on the mountain than others – and what about constituent named parts – South Col on Everest, or Striding Edge on Helvellyn – when are you on the one and when on the other, and and does the one become part of the main peak at some scale or distance of viewing?

For me advances in gazetteers will only come when they can answer the more subtle queries with appropriate maps, and that is only possible when some serious consideration is given to integrating the spatial footprint of the features with the gazetteer.

Workshop on Digital Gazetteer Research and Practice
Santa Barbara, California, 7-9 December 2006

National Geospatial-Intelligence Agency and United States Board on Geographic Names

Statement of Interest:
Dr. Beth Driver
Mr. Randall Flynn

The United States Board on Geographic Names (BGN) is the U.S. Government body that has been appointed, through act of Congress, as the national names authority for geographic names within the United States. It has also been given authority to set standard forms of foreign place names for use across the whole U.S. Government.

The National Geospatial-Intelligence Agency (NGA) is one of the sixteen agencies that make up the United States Intelligence Community (IC); since 2003 NGA has been continuing the critical national security missions of its predecessor organization, the National Imagery and Mapping Agency (NIMA).

NGA provides staff support to the BGN and maintains the Geographic Names Data Base (GNDB), which is the official repository of standard place name spellings for foreign place names for use throughout the U.S. Government. The GNDB currently holds more than five million place name entries for more than three million unique geographic features. Its coverage is worldwide, with the exception of the United States and Antarctica. Domestic features and geographical names are stored in a separate database called the Geographic Names Information System (GNIS) maintained by the BGN Domestic Names Committee and staff at the U.S. Geological Survey.

The GNDB was originally designed in support of military cartography in areas of the world where the United States had active conventional conflicts and generally covers scales of between 1:250,000 and 1:50,000. Recent changes in the nature of warfare towards asymmetric conflicts, as well as increased centrality of geospatial intelligence (GEOINT) within the IC, have forced corresponding alterations in the nature and content of the BGN/NGA gazetteer.

First, there has been a dramatic increase in the scale of information requirements from IC and DoD customers, from county-level to urban-level scales, and even to cadastral-level scales

in some areas. This demand for increased detail has led to challenges in collection, storage, and dissemination. An as-yet undetermined question is the proper amount of cross-references (variant entries) to add to the gazetteer: a delicate balance must be found between improved searchability and increased noise. Features generally found only on very large-scale maps have revealed previously unknown problems of competing names authorities; for example, discussions are under way between NGA's Office of Global Navigation and the BGN over who has ultimate authority to set standard names for lighthouses.

Increased focus on joint operations with multinational partners leads to the need for multilingual gazetteer services. To support this requirement, the GNDB became fully Unicode-compliant in 2005, and non-roman script names began to be added (about 300,000 non-roman script names to date). Enriched metadata has been found necessary to distinguish amongst the new classes of entries to the GNDB: language of the name, dialect, script, orthography, and transliteration system. Each of these domains of metadata suffers from challenges of definition; for example, does "language" refer to the etymology of a geographic name, or to the language in which it is currently used? Each new metadata domain as well poses difficult challenges of identifying and reconciling existing, and sometimes competing, standard enumerations.

Alterations to the GNDB's current name typing schema, which encodes name provenance and authoritativeness, have also been proposed, in response to the increased complexity of the entries in the enriched GNDB. A renewed interest in time-variant aspects of geographical names has led to lengthy discussions about how to encode historical change within the GNDB: at the feature level or the name level.

The explosion of third-party digital gazetteer data has revealed a major technological and resource bottleneck for maintaining the GNDB: how to conflate external gazetteers into the GNDB. Several independent research efforts are currently testing algorithms based on record-linkage methodology, which will give a measure of automation to the current manual conflation process. Beyond conflation, federated architectures and distributed gazetteers promise to ease the burden of maintaining gazetteer data on a global scale.

In conjunction with these changes, the BGN Gazetteer has been recognized as an important enterprise-wide service that needs to be available across NGA and the IC. This realization comes from

statistics that point out the critical role of place names in organizing information. For example, Microsoft and Yahoo recently reported that between 12 and 16% of all user queries on their general search sites contain at least one geographic name.[1] Similarly, a recent U.S. Government report notes that geographic location is a key feature of 80-90% of all government data.[2] The challenges and opportunities of geographic information retrieval (GIR) lead to questions of interoperability; more than ever there is a need for greater coordination between national and international standards organizations, such as the Open Geospatial Consortium, the International Standards Organization, and the Digital Geographic Information Working Group, for example.

The integration of the GNDB into the wider NGA enterprise platform has entailed another painful process: the transition in the gazetteer away from a strictly point-based geometry towards one where each entry may have an arbitrary geometry. Increased complexity of gazetteer geometry representations has meant increased cost of ingesting and maintaining locational information, and difficult choices are being made about costs and benefits of collecting various types of locational references for various features.

A final area of interest to highlight is the relationship of gazetteers to ontologies. What level of ontological information is most appropriate to store in a gazetteer? How can gazetteers be made more interoperable with the proliferation of geospatial ontologies? Current NGA/BGN efforts in this vein include the mapping of the GNDB feature catalogue to the NGA Feature Catalogue (NFC), as well as the federated linkage of the GNDB and several other NGA data stores to the Cycorp Upper Ontology.

Dr. Driver & Mr. Flynn look forward to discussing these and all the other topics of the Gazetteer workshop. Dr. Driver is the Chief of the University Outreach and Partnership Branch within NGA's InnoVision Directorate and is responsible for coordinating all NGA collaborations with the academic community. Mr. Flynn is the Executive Secretary for Foreign Names of the United States Board on Geographic Names, as well as the NGA Geographer, and is responsible for oversight of the BGN/NGA GEONet Names Server Gazetteer.

---

[1] SIGIR 2006, Workshop on Geographic Information Retrieval, Seattle, Washington, August 2006.

[2] Federal Geographic Data Committee, http://www.fgdc.gov/library/whitepapers-reports/white-papers/homeland-security-gis.

On Gazetteers

John R. Frank
MetaCarta Founder and CTO
November 2006

MetaCarta sells commercial geoparsing and geographic information retrieval systems. These products depend on data bundles called Geographic Data Modules (GDMs). Each GDM contains a gazetteer and a set of linguistic statistics that model the natural language usage of geographic references. These linguistic statistics capture nuances of human discourse. The MetaCarta GeoParser uses its GDMs to predict what phrases a human would extract from a document and resolve to particular locations. The MetaCarta Geographic Text Search (GTS) system indexes geographic and textual information from large collections of documents, so users can find everything known about any place.

Anyone can generate new geographic names simply by using them in communication. From the perspective of adding geographic structure to unstructured documents, all geographic references are valid. One has no recourse to enforce particular naming conventions. Authors' choices dictate. Thus, gazetteers are unbounded in scope.

While anyone can create new georefs, intuition suggests that a majority of literate people agree on a special subset of common locations. Plotting the frequencies of occurrence of georefs from a balanced corpus of data partially confirms this expectation. The distribution has a peak of high frequency names, but the distribution is fat-tailed. Most gazetteer efforts start in the peak and grow by expanding toward the tail.

Experience with peak frequency names underscores the value in the diverse swarm of locations beyond the peak. Plenty of structure remains outside the peak. For example, power law models might illuminate the social processes that bring a location into awareness.

Further segmentation is useful. MetaCarta organizes GDMs by language and genre of discourse. Frequency distributions vary across languages. Various genres use location names differently. For example, news articles often uses geographic names as metonyms for state actors. Short designators like "Building 4" or "Lease Block 22" appear frequently in corporate and technical discourse. Style and tone change the meaning of georefs in government reports, grade school texts, geology research, travel guides, and other genres.

This philosophical question deserves attention: What defines a location? Can we rely on a Platonic form for Location? If France, the Statue of Liberty, and the North Pole are to be instances of Location, then its definition must encompass many ideas at once.

Most uses of the word "location" imply a geospatial sense, and yet, geometry alone fails to capture the full meaning ascribed to a location. Simply listing polygon vertices

does not capture the essence of what people mean by locations like Aix-en-Provence or Rome or the Great Wall. One naturally thinks of history, political hierarchy, and other relationships that mankind has with that entity.

Is the oil platform known as "Jack" simply a longitude-latitude bounding box? Oil platforms' most striking features are their drills stretching into the depths and changing orientation during operation. Besides this geometric complexity, when such an entity enters the consciousness of more people, its name takes on new meaning. Just as Cazumel means "vacation" in a somewhat generic sense, the discovery of oil at Jack has taken on significance independent of its literal location – for some people.

In the process of communicating, people add attributes to location entities. As a location gains importance, communities wander away from asserting rigorous gazetteer definitions. Definitions escape. The line between defining and describing fades.

Locations often come into existence as bookkeeping conveniences. As they progress from mere gazetteer entries up the fat-tailed distribution of awareness, locations gain personality. Only natural language can capture the richness of these attributes.

What defines a location? The collective awareness of many people defines a location. Structured gazetteers cannot capture this – at least not without true artificial intelligence.

The emerging field of geographic information retrieval offers a pragmatic way forward. Instead of engineering more and more knowledge into gazetteers, we gain more ground by acknowledging the boundary where gazetteers stop and unstructured GIR takes over.

Gazetteers form a fundamental part of GIR. Articulating useful boundaries between these two types of tools will foster progress in both areas of effort. GIR offers a bridge toward cartography and information presentation, which connects gazetteer efforts to people whose awareness drives the organization of gazetteers.

Position paper: M. Freeston

Digital Gazetteer Workshop Santa Barbara, CA 7-9 Dec 2006

As a computer scientist with a long-held interest in database technologies, my interest in gazetteers stems from my work over the past twenty years on spatial database systems and georeferenced digital libraries. – notably the Alexandria digital library. For much of this time, and certainly the past ten years, it has been a puzzle to me why the potential power of search by georeference has not been a more widely rccognized and exploited paradigm.

In particular, it is at least at first sight surprising that neither the academic nor commercial database research community has so far developed adequate support for this paradigm. Within a database management system (DBMS), it is still hard if not impossible to represent the varying degrees of accuracy of georeferenced information, and there is certainly no direct support for an ellipsoidal frame of reference. Geospatial indexing and querying in database systems remains poor. The relative lack of commercial interest in this direction is however not hard to trace: a schism developed between the GIS and database communities as long ago as the 1960's, when the GIS community found that database technology was not able to support the basic functionality it required. In particular the GIS focus on geospatial visualization was not seen as relevant to the requirements of business database systems of the time. Nevertheless, as GIS moved from batch processing to real time, and the scale of GIS data management increased by orders of magnitude, GIS systems developed 'loose couplings' to DBMSs. Even so, at the conceptual level at least, the representation, querying and visualization of geospatial objects remained – and remain today - a function of the GIS system rather than the DBMS. For example, the components of the representation of a spatial extent in ArcInfo can be stored in a loosely-coupled DBMS, and search operations on such extents utilize the standard database index methods (of which, until only ten years ago, none were inherently spatial). But the interpretation and manipulation of these components as parts of a single spatial object, and its visualization, remain within the GIS system.

The complete integration of GIS and DBMS therefore remains an unsolved problem. But the world has moved on, and now the problem has become much more challenging. Two particular aspects interest me:

1) the use of ontologies to express not only a wide variety of geospatial data types, but relationships between them;
2) the representation, manipulation and querying of geospatial data on a global scale;

It is now widely recognized that there is a need for the creation of an exhaustive domain ontology in almost every field of knowledge. In the case of geospatial information, the humble gazetteer has now grown into a fully-fledged ontological framework. Much work has already been done on defining a gazetteer content standard and access protocol. The challenge now is to integrate this framework, and a reasoning engine over this framework, into a DBMS.

However, this cannot be satisfactorily achieved without addressing the fundamental problem of how to represent and manipulate geospatial data within a DBMS. I am particularly interested in the development of a geospatial partitioning and addressing mechanism based on the recursive division of the surface of the Earth into a hexagonal grid. The objective is to find a way – or at least a satisfactory compromise between conflicting requirements – of representing geospatial extents accurately at different scales.

Rethinking Gazetteers and Interoperability

Greg Janée
Institute for Computational Earth System Science
University of California at Santa Barbara
Santa Barbara, CA 93106-3060

December 9, 2006

1 ADL gazetteer protocol

The Alexandria project's desire to create a gazetteer protocol grew out of two needs. First, we had identified gazetteers as a key component of our architecture for distributed geospatial digital libraries, but having only one instance of a gazetteer to go by (ours), it was unclear which features digital libraries could generally expect from gazetteers, and which were artifacts of our implementation. Second, our initial work on gazetteers was largely content-based, having focused on developing the ADL Gazetteer Content Standard (GCS) [1], and a protocol would provide a complementary, functional definition of a gazetteer as a type of knowledge organization system.

Thus in 2001 we collaborated with ESRI to create the ADL Gazetteer Protocol [2]. In the interest of interoperability, the protocol defines a simplified gazetteer model that is compatible with the GCS, but omits much of the latter's elaborate details such as temporal qualification and lineage. In the simplified model, gazetteer entries are equated with conceptual places; each entry (i.e., place) is described by one or more names (notably, all names are unqualified), one or more spatial footprints, and zero or more feature types. Each of the aforementioned quantities may be flagged as current or historical, and one quantity within each category must be flagged as primary. The place as a whole may be flagged as current or historical. Named relationships may be asserted between places, though the relationships themselves are undefined by the protocol. The protocol defines just two services: download all entries, and search for entries that satisfy a boolean combination of constraints on place attributes.

Our experience with the protocol has generally been favorable. It fulfilled our original needs, and we and a few others have successfully implemented it on both the client and server side. However, several limitations have emerged, notably:

- *Lack of support for qualified placenames.* Searching for a place qualified by the name of a containing (and disambiguating) administrative unit is not directly supported by the protocol, and even assembling such searches out of the lower-level functionality the protocol *does* provide is onerous at best. This is a significant limitation given the ubiquity of such queries.
- *Conflicting and unpredictable semantics between spatial and relational searches.* The ADL protocol provides two ways of expressing containment constraints: searching spatially (e.g., find "Santa Barbara" spatially contained within California's footprint) and searching relationally (find "Santa Barbara" that has a PartOf relationship to California). Conceptually, these two types of queries are equivalent (or, at any rate, any difference between them is surely splitting semantic hairs), yet the results are likely to be entirely different due to implementation artifacts such as incompletenesses in the recording of

PartOf relationships, and problems testing spatial containment due to spatial footprints being overly small (points) or overly large (bounding boxes). Furthermore, these differences are, from the user's perspective, both unpredictable and variable from query to query.

The OGC's WFS gazetteer profile [3] is analogous to the ADL protocol. Though there are a number of significant differences, the two protocols are largely similar in their broad characteristics, and they target the same level of functionality. As a consequence, many of the statements regarding the ADL protocol below apply to OGC's as well.

2 Interoperability use cases

As time has passed since the initial development of the ADL protocol, and as mapping and geocoding services and "mashups" have become more common and easier to implement, we have found ourselves stepping back and asking questions such as: What does interoperability mean for gazetteers? What role can gazetteers play in some of these new services? Does it even make sense to speak of a *gazetteer* protocol? To begin to answer these questions, let's observe five use cases related to gazetteers.

Harvest Retrieve the entire contents of a gazetteer. Harvesting is necessary to support aggregation of gazetteers, particularly if we want to provide unified search over multiple, local, idiosyncratic gazetteers. Harvesting is also necessary to support certain intensive uses of gazetteers, such as in geoparsing, where any online protocol is bound to be too inefficient. Harvesting requires: a standard means of retrieving content, such as OAI-PMH [4], and if not a common representation of gazetteer content, at least a few well-known representations. A common feature typology would be helpful, but this is probably destined to remain a pipe dream.

Lookup Find a place by name or other description. This is classic gazetteer functionality, of course, but with this use case we're explicitly looking beyond gazetteers and including, for example, street address geocoding. Indeed, the trend in recent mapping services (Google, Yahoo, etc.) is to recognize more and more forms of spatial reference (airport codes is a good example), purely for user convenience. Note also that these services desire only simple point locations in return, in order to define a location of interest or a place to pan a map to; there is no use of extended placename information in this use case.

Reverse lookup More classic gazetteer functionality: find places near a given spatial location. More specifically, find places of a given type near a given spatial location. And even more specifically and possible more useful in many circumstances: find the *nearest* place of a given type to a given spatial location. As a hypothetical example of a use of the last type of query, consider a service that offers airport-related information. Users of the service would like to enter the most convenient (to them) form of spatial reference to identify an intended airport. Meanwhile, the developer of the service would like to take advantage of a lookup service that translates spatial references to airport codes, for the latter is how the service's information is likely to be indexed. A lookup service, such as described above, provides half the solution by returning latitude/longitude coordinates; a reverse lookup service can then be used to return the nearest airport.

Geoparse Identify and geocode the placenames in a document. This use case involves gazetteers, but is, strictly speaking, outside the scope of gazetteers for two reasons. First, geoparsing requires examining substantial context around candidate placename references, often the entire document. Second, a geoparser's use of a gazetteer is computationally intensive and specialized, and hence the gazetteer will need to be held close to the geoparser, say, in a local database. A nice example of a "mashup" service that combines geoparsing with another, widely-used protocol is the GeoNames.org RSS-to-GeoRSS converter [5, 6, 7], which automat-

ically geoparses and geocodes articles from any RSS feed, thereby allowing the articles to be displayed by a GeoRSS-capable map service such as the Acme GeoRSS Map Viewer [8].

Ontology Along the lines of the Semantic Web, we can think of a gazetteer as a (potentially distributed) kind of knowledge base of facts and associations that supports certain kinds of inferences. The key requirements here are that each concept (i.e., place) be uniquely identified (in RDF, this means by URI) and that associations be defined by an ontology. However, significant barriers remain to implementing this use case. If we are to work with more than one gazetteer at a time, then either there can be only one fact for each place, or unification of equivalent facts (i.e., places) is necessary; inferencing will fail otherwise.

3 ADL protocol, revisited

Returning to the ADL gazetteer protocol, let's look at how well it supports the above use cases.

Harvest The ADL protocol's simplified gazetteer model may provide a useful, "lightweight" counterpart to the GCS. Otherwise, the protocol's harvest service, lacking a restart or resume capability, is too simplistic to support large-scale downloads. In any case, OAI-PMH is well-established in this area.

Lookup ADL's lack of support for qualified placenames, and generally rigid query model, are too limiting.

Reverse lookup Reverse lookups are supported, but not nearest place queries.

Geoparse, Ontology These are outside the scope of the ADL protocol.

Thus it does not appear that a gazetteer protocol is, by itself, all that supportive of use cases that involve gazetteers. In the harvest use case, gazetteer interoperability centers more around representation; a general harvest protocol satisfies the functional needs. In the lookup use case, a protocol that is generalized beyond gazetteers is required. In the geoparse use case, gazetteers play a critical role, but not in an online sense; again, interoperability is centered around representation and harvesting. And likewise for the ontology use case.

4 Question for the workshop

As stated earlier, the ADL gazetteer protocol has proven useful, and will continue to be useful for accessing and defining gazetteer functionality. But in thinking of future development directions, should we rethink our approach to gazetteer interoperability?

In our past work we started with an entity (a gazetteer) and then defined a protocol giving access to that entity. Should we instead orient our efforts along the lines of the aforementioned use cases, and define protocols that are oriented around *functionality* (lookup, reverse lookup, etc.) as opposed to *entities*, protocols that various kinds of entities (gazetteers, geocoding services, etc.) can implement to varying degrees?

Simply refactoring functionality is not going to solve any difficult problems, of course: spatial search semantics will continue to be problematic and messy. But the suggested refactoring can provide value in a couple ways:

- The resulting protocols are likely to be more closely aligned with user needs and desires.
- The distinction between functionality (geocoding, say) and the entity implementing that functionality (a gazetteer or geocoding service or other type of entity) is clarified.

References

[1] Linda L. Hill (2004). Guide to the ADL Gazetteer Content Standard, version 3.2. `http://www.alexandria.ucsb.edu/gazetteer/ContentStandard/version3.2/GCS3.2-guide.htm`

[2] Greg Janée and Linda L. Hill (2001). The ADL Gazetteer Protocol.
http://www.alexandria.ucsb.edu/gazetteer/protocol/

[3] Jens Fitzke and Rob Atkinson, eds. (2006). Gazetteer Service — Application Profile of the Web Feature Service Implementation Specification. OGC 05-035r2, version 0.9.3.
http://portal.opengeospatial.org/files/?artifact_id=15529

[4] Carl Lagoze and Herbert Van de Sompel. The Open Archives Initiative Protocol for Metadata Harvesting.
http://www.openarchives.org/OAI/openarchivesprotocol.html

[5] Really Simply Syndication.
http://en.wikipedia.org/wiki/RSS_(file_format)

[6] Geographically Encoded Objects for RSS feeds.
http://www.georss.org/

[7] GeoNames RSS to GeoRSS Converter.
http://www.geonames.org/rss-to-georss-converter.html

[8] Jef Poskanzer. ACME GeoRSS Map Viewer.
http://www.acme.com/GeoRSS/about.html

Biography

Greg is a research specialist for the Institute for Computational Earth System Science and for the Map & Imagery Laboratory, Davidson Library, at the University of California at Santa Barbara. He is currently working in two areas: methods and architectures for discovery of distributed, heterogeneous geospatial data and, more generally, digital library support of Earth science data lifecycles; and long-term preservation of geospatial data. The latter work is for the National Geospatial Digital Archive.

Previously, Greg was technical leader of the Alexandria Digital Library Project, principal developer of the Alexandria Digital Library software, and developer of the ADL gazetteer and thesaurus protocols.

His experience prior to the Alexandria project was in software engineering for commercial and government clients in the areas of object-oriented class libraries; 2D, 3D and fractal-based visualization; rule-based expert systems; compilers and query languages; and embedded database systems.

Greg holds an M.S. in computer science and a B.S. (summa cum laude) in mathematics, both from the University of California at Santa Barbara.

Greg's publications and conference and workshop presentations are listed at http://www.alexandria.ucsb.edu/~gjanee/publications.html.

Statement of interest for the Digital Gazetteer Research and Practice Workshop, December 2006, by Gregory B. Hill, University of Colorado.

My interest in gazetteers stems from recent experiences developing web based spatially enabled gazetteers and applications that depend upon them. The gazetteers themselves are distributed instances with a common schema, inspired by the Alexandria Digital Library Gazetteer Content Standard[1], and developed as part of the Biogeomancer[2] suite of georeferencing tools for natural history and biodiversity. The main client application is called GRIPPER[3]. One of the distinguishing features of GRIPPER is its use of a gazetteer to catalogue tiled raster layers. Via this means, gazetteer queries can be enhanced through the application of filters based upon environmental conditions or species occurrence data. By way of example, a query upon wetlands within Labrador could be further refined by specifying an average temperature range, or a species identifier. GRIPPER provides a flexible and powerful query building interface[4] and the ability to modify feature data in a user-specific gazetteer instance. It also allows features returned to be added to a selection layer; selection layers in turn are visualized using an integrated map portal[5].

The following paragraphs describe in more detail the work I have been doing. The workshop will be an excellent opportunity for me to present these developments to others working in the field and to learn from their questions and comments as well as to interact with researchers and developers who have been working and thinking about related challenges.

Gazetteer data has been ingested from various sources and normalized to WGS-84. The placenames are run through automatic character set detection during ingest and converted to UTF-8.

.

Current datasets ingested include the Getty Thesaurus of Geographic Names (TGN)[6], Geonet Names Server (GNS)[7], Geographic Names Information Service (GNIS)[8], U.S. Roads, Worldwide Protected Areas (reserves), and Worldwide Administrative Boundaries. The feature count exceeds 100 million, and will likely exceed 500 million in the future. Most features include MULTILINE or MULTIPOLYGON spatial descriptions. There is no hard limit on the maximum number of features; a distributed design with multiple gazetteer instances keeps the maximum virtual gazetteer size open ended and accessible. Each instance corresponds to a dataset and can reside on an arbitrary server.

---

[1] http://www.alexandria.ucsb.edu/gazetteer/ContentStandard/version3.2/version3.2.html

[2] http://www.biogeomancer.org/

[3] Geographic Referencing for Investigating Phylogenetics and Pathogenicity. R&D demonstration build running at http://biogeobox.colorado.edu:8080/gripper, please enquire for demonstration.

[4] For example, queries can be run in batch mode, and counts for the number of results can be obtained prior to running the queries. For queries returning many results, it is possible to iterate through result sets for a given query. Queries can be held in memory, modified, and re-run. Results are displayed in a UI incorporating recent advances in asynchronous technologies such as AJAX.

[5] Currently using the deegree WFS/WMS/WCS portal, see www.deegree.org

[6] http://www.getty.edu/research/conducting_research/vocabularies/tgn/

[7] http://earth-info.nga.mil/gns/html/index.html

[8] http://nhd.usgs.gov/gnis.html (not all of this data has been processed at this time).

Coordinated use of different classification schemes is handled via hierarchical relationships that are captured in a simple database table that is transparent to the gazetteer server code. For example, if a TGN classification term bears a child relationship to an ADL classification term, any search for the ADL term will also return features matching the TGN term. Similarly, if an ADL term bears a child relationship to a TGN term, searches on the TGN term will also return features classified using the ADL term. Equivalency is indicated in a similar fashion. Searching of equivalent and child terms can be turned on or off. This does not resolve all issues involved with multiple classification schemes, but it is backwards compatible: a dataset developed with a given classification scheme can still be searched as before using the assigned terms.

The geometric content is comprised of 0, 1, or 2 dimensional representations, a radius calculation based upon the concept of 'spatial fit'[9], a GML 3.1[10] representation, and several other forms designed to be used as data sources for the map portal. The taxonomic and raster extensions do not affect regular gazetteer content. For example, the raster filter capability is implemented not by modifying the content standard, but by adding raster tiles as features to a special gazetteer instance.

When queries are run, results can be edited. Classifications, alternate names, and the display name can be modified, added, or removed. The geometric representation can be translated. When these edits are committed, they go into the user's local database, not to the released databases used for searching. A secure, administrative logon-based procedure is used to commit modified records from user databases to the shared released databases, but users can immediately take advantage of the results of modifications to their own gazetteer instance. In the ordered list of search databases for each user, the user's gazetteer instance is first in the list, so their edits are returned in the first set of matches.

This open source software architecture is primarily based upon AJAX[11] (zk[12]), JSF[13], Struts[14], Java, deegree (web map client supporting WMS/WFS/WCS), PostGIS[15], and other leading open source technologies.

[9] http://bgdev.berkeley.edu/?q=node/568
[10] http://www.opengis.net/gml/
[11] Asynchronous Java XML
[12] zk1.sourceforge.net
[13] Java Server Faces: jsf.sun.com
[14] Struts.apache.org
[15] postgis.refractions.net

Digital Gazetteer Research and Practice: Position Paper

Bruce M Gittings, University of Edinburgh

Scotland and Gazetteers

> "Next to a good dictionary, the most generally useful book is a good gazetteer"
> -- W.G. Blackie (1855)

Through the 19th Century Scotland maintained a strong tradition of the production, and use, of quality gazetteers. These volumes described the geography of Scotland, Britain and indeed the World. In terms of Scotland itself, the culmination of the art was undoubtedly the *Ordnance Gazetteer of Scotland* compiled in 1885 by Francis Groome, which still remains the standard geographic reference text to be found in major libraries on Scotland. This work was updated in the 1890s to reflect significant administrative changes, with a final edition published posthumously in 1903. Since then, the *descriptive gazetteer* has largely been replaced by the tourist guide and the place-name list (*short-form gazetteer*), the latter being the form which is most familiar to those of us involved in geographical information science (GIS). The tourist guide is no replacement however, in that they often introduce unacceptable bias, either in terms of selection of places described or what is actually said. Yet, with likes of Google Maps and Microsoft local.live are finding that basic mapping and vertical aerial photography have their limits and are clamouring for information. Thus, the need for a richer geographical description (through imagery, text and enriched databases) is becoming a focus of attention. Equally a broader community interests in heritage, local history and genealogy are developing apace. The importance of family history tourism to the Scottish economy is becoming significant, reflected by government resources being out into initiatives such as Scotland's People (www.scotlandspeople.gov.uk). Also initiatives such as the Millennium-lottery funded Scottish Cultural Archive Network (SCRAN) has given rise to the digitisation of a range of heritage resources, including papers, artefacts, art-works, books and maps held in a range of private and public collections. What connects all of these together is geography and place. Yet, in the United Kingdom, we have no authoritative source for place-names information and, worse, librarians and others without a geographical understanding are beginning to propose 'solutions' which range from the unworkable to the bizarre.

The Gazetteer for Scotland Project

Thus, while high quality geographic information in the form of maps and images have gained widespread acknowledgement as an important resource, descriptive information is becoming rare - largely because it needs expert collation, writing and editing - and texts on Scotland were becoming increasingly out-of-date. The only topographical directories for Scotland produced this Century have been place-name lists published, for example, by the Ordnance Survey and the Registrar General for Scotland and the Johnston's Gazetteer of Scotland (last updated 1973) which is limited in content and now out of print.

Against this background, David Munro (now Director of the Royal Scottish Geographical Society) and I created *The Gazetteer for Scotland* in 1995. Initially intended to be published as a book, the project was soon directed towards the Web, which was a new medium at the time. This decision has had a number of advantages. As the magnitude of the project became clear it was increasingly obvious that the amount of information produced would be well beyond the limits imposed by a modern-day publisher. Equally, the web permitted the incorporation of a large number of georeferenced photographs, allowed draft text to be modified following exposure to a critical public and the inclusion of a variety of interactive facilities which make the information more accessible. Thus the Gazetteer for Scotland has been built as a database based on Oracle, making use of its

relational features to create a web of interconnections between individual entries which are reflected on the web site. The database was extended well beyond settlements and geographical features, to include biographies of famous individuals and descriptions of families which recur through the places listed. The database now extends to 13400 entries, 1.2 millions words and 6500 photographs and has been built through novel geographical research as well as secondary sources. It has been further extended through the incorporation of the six-volumes of Groome's 1885 work to give historical context and now approaches 1 million hits per week. This work has given rise to considerable reflection on the changing face of Scotland over the last 120 years as well as a unique insight into the difficulties of connecting two quite different gazetteers using a combination of computer-matching algorithms and manual methods.

A Place Name Authority for Scotland

Within the United Kingdom, we have no Place Name Authority. We do have a body called the Permanent Committee for Geographic Names, sponsored by the Ministry of Defence and who liaise closely with the United Nations. However, perhaps in deference to our imperial past, this body looks at everyone else's place-names, explicitly not considering place-names in Britain. Along with many other countries, Scotland is in the process of building a SDI in the context of a GI Strategy which was published a year ago and is currently influencing an over-arching United Kingdom GI Strategy. Since 1997, Scotland has had a devolved government (within the UK) with broad powers over many areas and this has presented certain opportunities. Within this strategy, a National Place Names Gazetteer has been proposed, following considerable lobbying of government by myself. This will complement a Digital National Address Gazetteer (DNA), perhaps peculiarly, a project which is already well advanced and close to launch, encouraged by national security fears. The concept of a National Place Names Gazetteer is, however, not well advanced and, indeed, deprecated by some despite the United Nations encouraging all countries to have a place-name authority in place (Resolution IV of the UN Conference of Geographic Names). Thus, while postal towns will be incorporated within the DNA, this will deal with neither communities, nor non-addressable geographical features. All of these represent a significant problem in Scotland, where it is not at all unusual to see a different rendering of a name on the map, from the road sign on the way into a settlement and from that used on the village shop or post office. This confusion is further exacerbated by the policy of certainly local authorities, and now the ordnance Survey, to adopt a policy on Gaelic names, which involves the translation of English names into Gaelic in those areas where Gaelic is still spoken. Even in the traditionally English-speaking parts of Scotland, it is not unusual to have the same place-name (river name or mountain name) repeating within a relatively small area. These issues make it extraordinarily difficult to distinguish one place from another. This is reported as being an issue for the emergency services amongst others, especially in a climate which means that emergency calls are now taken by a few large call centres, rather than at individual telephone exchanges where the local knowledge of the operator had value.

It is vital that not only is a Place Name Authority created, but that this gives rise to a free-to-use definitive interoperable gazetteer service that is properly resourced and maintained.

The Dangers of Non-Authoritative Sources

While there is a view that the Wikipedia model will solve all of these problems, I have significant concerns. Putting aside my seemingly constant battles against the plagiaristic tendencies of Wikipedia contributors, which I fear will discourage professional, fully-researched contributions and hence the value of the resource, experience has suggested it likely that a publicly-contributed database will be spatially incomplete, inaccurate, and often considerably biased. In Scotland, place can have political connotations, with those of a Nationalist disposition having a very different

interpretation of place based on historical events. To allow public input / revision is one thing, but at some point the professionals need to take over.

Workshop on Digital Gazetteer Research and Practice
Santa Barbara, CA – December 7-10, 2006
Position Paper

Stephen C. Guptill
U.S. Geological Survey

Background

The U.S. Geological Survey developed the Geographic Names Information System (GNIS) for the U.S. Board on Geographic Names as the official repository of domestic geographic names data. Confusion and controversy about geographic names and their applications to places and features led President Benjamin Harrison to establish the U.S. Board on Geographic Names in 1890. That early Executive Order was based on recognition that conflicts in naming geographic features were, in fact, a serious detriment to the orderly process of exploring and settling this country. A later decision, in 1906, by President Theodore Roosevelt to extend the responsibilities of the Board to include standardization of all geographic names for Federal use was a far-reaching decision that, coupled with the Harrison order, forms the foundation for the present organization of the U.S. Board on Geographic Names established in Public Law 80-242, signed by President Truman in 1947.

GNIS is the official vehicle for geographic names use by all departments of the Federal Government and the source for applying geographic names to Federal electronic and printed products. GNIS can be accessed at http://geonames.usgs.gov. The GNIS contains information about physical and cultural geographic features of all types in the United States, associated areas, and Antarctica, current and historical, but not including roads and highways (under statute, the Board on Geographic Names has purview over road and highway names, but has chosen not to execute that authority). The database holds the Federally recognized name of each feature and defines the feature location by state, county, USGS topographic map, and geographic coordinates. Over 1.9 million features are represented in GNIS. Other attributes include names or spellings other than the official name, feature designations, feature classification, historical and descriptive information, and for some categories the geometric boundaries.

Components of Gazetteer Services

Geospatial Location:

Most features in GNIS are associated with one location (e.g. summit, structure, lake, populated place) even though that feature may have an areal extent. Streams have a location for the source and mouth. The extent of an areal feature can be estimated from a set of coordinates that identify a series of geographic tiles with one point per USGS topographic map containing the feature. More accurate geographic descriptions of the features are not maintained within the gazetteer, but are contained in thematic geodatabases. For example, the National Hydrography Dataset (NHD, http://nhd.usgs.gov), which is the USGS's most comprehensive set of digital spatial data about surface water features, is closely integrated with GNIS. The hydrographic feature names contained in and displayed by the NHD are from the GNIS. When partners submit new data to NHD, feature names are validated against the GNIS. To meet legal and policy requirements of the Board on

Geographic Names, the NHD will display a hydrographic feature name only if the feature is entered into GNIS with an assigned Feature ID. Similar linkages exist for administrative boundaries, where accurate geographic descriptions are maintained by the Census Bureau.

Efforts to delineate of precise extents for ambiguous physiographic features (e.g. summit, range, valley) have had limited success or utility. For example, it does not seem useful to tag each cell of an elevation model with a set of feature applicable names (Mount Rag, Blue Ridge Mountains). A way to deal with these soft-edged features is awaited.

Integration of Gazetteer Data from Multiple Sources:

The 30-year GNIS data compilation program began in 1976 and is continuing. The first phase (1976-1982) collected names (except roads and highways) from the USGS topographic maps, but many manmade and administrative features either are not shown or not named on these maps. Between 1982 and 1984, names from other Federal sources were collected, but only about 30 percent of the known names appeared on Federal sources (for manmade features it was a far smaller percentage). A second extensive compilation phase was begun in 1982 and continues to collect, State by State, data from official State and local sources as well as from other pertinent current and historical materials. Feature additions or corrections are accepted for consideration from any source, and when validated by appropriate agencies, will be entered into the database. Local and State agencies are encouraged to submit data and to participate in the GNIS partnership program. Non-government organizations with valuable data are considered on a case basis. Authorized partners have access to web based transaction entry and edit forms, which submit data directly to the GNIS for review and inclusion in the database. Partners also submit batch files in most standard formats, and coordinate with the Geographic Names Project to develop joint services, processes, and applications for greatest efficiency. Data entered into the GNIS immediately is available to all web services and applications dependent on it. While we anticipate discovering most additional entries (even historical locations) through the partnership program, there will always be those that escape detection. Web-based facilities allow users to submit an administrative name or errors they believe to have found. Because of the resources that would be required for official names validation, we have not yet implemented a “geoname Wiki” to solicit additional names data.

Interoperable Gazetteer Services

GNIS assigns a unique, permanent feature identifier, the Feature ID, as the only standard Federal key for accessing, integrating, or reconciling feature data from multiple data sets. The GNIS collects data from a broad program of partnerships and provides data to all levels of government, to the public, and to numerous applications through a web query site, web map and feature services (ArcIMS map service, GNIS XML service), file download services, and customized files upon request. GNIS has active linkages to a variety of mapping services including *The National Map*, Google Map, TerraFly.com, TopoZone.com, TerraServer, Tiger Map Server, and EPA’s Find the Watershed.

Lee Hancock
Founder and CEO
go2 Directory Systems

Digital Gazetteer Research and Practice Workshop, December 7-9, 2006

In 1994 it became apparent to me that technology was going to cause a monumental increase in devices and systems that would be able to precisely determine location and to create, store, organize and provide access to incredible quantities of local information. It also occurred to me that location determining capabilities and devices (primarily GPS) would eventually be available to the masses but that their usability and adoption could be enhanced greatly by the development of new and user friendly geo-referencing and user interfaces.

My initial concerns focused on ensuring that a variety of consumer devices, including cellular phones, watches, etc., were optimized with 'dial tone' friendly, simple and easy user interfaces and referencing systems. It soon became apparent, however, that the acquisition, storage, organization, manipulation and dissemination of accurate and comprehensive local information were also problems in desperate need of addressing. The analogy I have often used is that while GPS and other technologies were clearly creating new and very powerful "engines" that could revolutionize local information capabilities for the masses, inadequate efforts were being made to create a corresponding increase in the "fuel" for these new and revolutionary engines. Accordingly, the new GPS and other location technologies were in desperate need of new systems and interfaces to fuel these new capabilities and allow them to more quickly achieve their full potential.

The original efforts in 1994 resulted in the filing of an initial patent in August, 1996, significant efforts to educate and evangelize to the GPS and GIS communities to address the issue in from 1998 to 2002, and the launch of go2®, the world's first location-based directory available over mobile phones, in 1999. This latter product initiative and go2's technology and business model ultimately achieved broad distribution across virtually all major wireless carriers in the U.S. (AT&T, Cingular, Nextel, Sprint, and Verizon) and attracted the attention of significant industry resources from SAIC, Verisign, and Amdocs as strategic partners and investors. Like many high technology companies started inside the dot-com bubble, however, go2 was unable to sustain its growth and had to retrench significantly during 2002.

Notwithstanding significant downsizings and various legal restructurings, go2's mobile, local search and directory applications and services managed to survive to cross the proverbial market adoption 'chasm' to better times in 2005. In 2006, go2 was recently

ranked as the 5th most used Mobile Search engine behind industry giants Google, Yahoo, MSN and AOL, and since go2 is the only one of those mobile search engines that focuses exclusively on local search, there is a strong likelihood that go2 is the leading LOCAL, mobile search and directory service in the U.S. Since go2's launch in 1999 go2 has delivered over 1 Billion page views of mobile, local search information and content to tens of millions of unique mobile phone users. go2 has logged a great deal of information about each of those page views, including the specific search request (typically category or specific business name), date, time, carrier, device, location (either automatically determined or selected by end users through a variety of systems). This information provides a wealth of knowledge about when and where end users are and what they are looking for when they are using their mobile phones for local search and information.

go2's significant experience delivering local information through handheld devices lends credence to its view that there is still a significant need for better user interfaces and more accurate, deep, timely and compelling local information. While today go2 is focused primarily on the delivery of consumer oriented information – beginning with the location and information related to consumer destinations and businesses – we believe that successful partnerships between private and public sector organizations will ultimately expand beyond basic consumer information and enable a plethora of compelling local content to be created, acquired, stored, and made readily and easily available for the masses.

There are a number of benefits to achieving these objectives well beyond the obvious utility of obtaining accurate information anytime and anywhere. These benefits range from the ability to increase awareness, access and use of educational, historic, and other information (e.g. environmental information) to the ability to utilize an ecosystem of local search activity to create information and better understand all types of population movements and needs. While some of these are likely to create commercial opportunities and generate commercial benefits, we submit that establishing the right ecosystem and cooperative efforts between public and private sector organizations will ultimately accelerate the development and adoption of valuable local information systems and capabilities. These new partnerships and systems will benefit numerous participants in both the public and private sectors to the ultimate benefit of the general public.

November 16, 2006

JORDAN T. HASTINGS

Position Statement
DGRP Workshop

I regard a digital gazetteer (DG) as the quintessential geographic information system (GIS) having just one kind of feature, place, with two required descriptive attributes, a free-text placename and a categorical placetype. The geospatial attribute, footprint, is interpreted directly by the GIS. Despite this apparent simplicity, a DG pushes GIS technology in many ways: footprints may be approximate, placetypes are often fuzzy, and placenames have important linguistic connotations. All three descriptors are time-dependent.

Moreover, a single place may be described by multiple placenames, placetypes and/or footprints concurrently. A critical – perhaps the critical – task in gazetteer construction, therefore, is place *identification*, i.e. determining when two (or more) different descriptions in fact apply to the same place, and conversely when the same descriptions apply to different places. Bigler, Daowaga, Tula Tulia, and Lake Tahoe are all names for the same place, whereas Lake Geneva might be a city in New York, or in Wisconsin, or a water body in Switzerland. Place identification also presents itself after the fact, in that need to detect and remove entries for "duplicate" places from an extant gazetteer, or alternatively to conflate them into a single entry. Finally, place identification is essential to federating and ranking results from distributed gazetteer queries.

Objects with indeterminate boundaries, differing names, fuzzy types, time dependencies, etc. appear in GIS applications other than gazetteers, of course. Coming to grips with these issues in the context of DGs is appealing for two reasons: 1) because of their relative simplicity as GIS, gazetteers cameo the underlying knowledge engineering and data management issues, without loss of generality; and 2) gazetteers are increasingly utilized in data mining, information retrieval and Web query applications, so correct place identification and conflation have practical importance.

[end]

Extending Gazetteer Service in Geographic Information Retrieval

A Position Paper

Naicong Li, Jordan Henk
The Redlands Institute, University of Redlands

Hill (2000) identifies three basic elements in a digital gazetteer: *name*, *footprint*, and *type* (or category). Gazetteer entry *types* are organized in a hierarchical classification system. The name and footprint properties of geographical features enable automatic association between non-spatial data (e.g. text documents) and the location of geographical features, whereas the feature type hierarchy provides automatic organization of the the spatially indexed data. Used in a geographic information retrieval system (GIR, Larson 1996), such spatial indexing makes it possible to answer thematic ("what") and locational ("where") queries. The "theme" in the query may be spatial, referring to a geographic feature (e.g. the City of Redlands) or to a category of such features (cities). It can also be non-spatial, referring to some property of the feature (e.g. the city's population). While the *name* and *footprint* properties of geographic features are formally defined in a gazetteer, the typical set of properties associated with a particular feature category often are not. Similarly, the typical set of relationships among geographic features (other than the *is-a* relationship, such as the *capital-of* relationship) are often not present in gazetteers. GIR systems could potentially benefit from explicit definitions of these properties and relations by using them to automatically index and later retrieve relevant data, especially in the case of highly structured data such as GIS datasets.

The Redlands Institute at the University of Redlands is conducting applied research in GIR using a prototype platform known as the Geospatially Referenced Information Portal (GRIP). GRIP was designed to facilitate knowledge synthesis across a disparate set of information resources (sources such as internet engines, local file systems, digital libraries, databases, etc. - all stored in heterogeneous formats, possibly with non-standardized structures). GRIP (which is being extended to support knowledge management tasks for a desert tortoise science and habitat management project) currently uses a project gazetteer with a feature type hierarchy modeled after the ADL Feature Type Thesaurus (FTT). This gazetteer has over 12,000 features extracted from recent versions of the USGS GNIS gazetteer (for the states of CA, NV, AZ, and UT); USGS Federal Land polygons augmented with boundaries for states, counties, state parks, etc.; and further supplemented with project-based GIS data. Using this gazetteer, GRIP supports basic spatial searches and basic queries (including theme and location).

Our users are often interested in the exact value or range of values of a certain property of a geographic feature (e.g. the actual tortoise population density of a critical habitat unit), or the statistics on these values, or retrieving a subset of geographic features of a certain type based on some specification of their property values. Sometimes the data value already exists in the project GIS dataset(s), sometimes GRIP needs to perform spatial and/or non-spatial processing to derive the answer. To handle such queries, we have adopted an approach similar to those described in Fonseca et al. 2002 and Lutz et al. 2006. We are prototyping using a light weight (not heavily axiomatized) project ontology for our user community. This ontology contains project-based concept categories (both spatial and non-spatial). For each concept category, we define a set of relevant properties (e.g. the 'area' of a 'habitat'), including definitions of the property value type and value range. We also define its relationship (of types *is-a*, *partitive*, *causal*, *functional*, *associative*, etc.) to objects of other categories when applicable. Data in the project database are linked to appropriate categories as instances of these categories. Such links are established by a process similar to "registration mapping" as described in Owers and Ludäscher 2004, and in Lutz and Klien 2006. The categories in this ontology thus can be used as an index into the data in the project geodatabase. The categories can also be used to index document metadata, and they may be used to annotate text documents (through recognition of their corresponding natural language expressions in the documents) to produce sub-document level metadata for various levels of discourse units such as word, phrase, sentence, paragraph, etc. This ontology is designed to import pre-established upper level ontologies including an ontology of spatial relations with definitions of topological relations, distance and direction. It also incorporates definitions of frequently used spatial and non-spatial operations which can be invoked when GRIP is processing the more complex spatial search

(Larson 1996, Jones et al. 2004). Using the properties, relations and rules defined in this ontology, GRIP will be able to deduce implied information through inferencing and reasoning.

GRIP exposes this ontology to its user in one of its query formulation interfaces. One advantage of ontology-driven query formulation is that the category, instance and property in the ontology may be used as a set of controlled vocabulary terms in query formulation (thus by-passing the need of disambiguating user queries in natural language). Another advantage is that this approach facilitates user exploration (and therefore discovery) of the information stored in the project database(s). A novice user, unfamiliar with desert tortoise habitats, may browse the ontology and learn about relevant factors (such as climate conditions, vegetation cover, soil, threats, etc.) through the property definitions for the concept of "Desert Tortoise Habitat". The user may also obtain a more comprehensive knowledge of threat factors by browsing all the sub categories of "Threat".

Our research is intended to produce a generic platform in which a project ontology and a project gazetteer may be integrated to enhance geographic information retrieval. Because of its relatively simple structure and the use of an efficient indexing scheme, the gazetteer has a much faster performance time during a search. The gazetteer is also sufficient for answering the most common types of queries regarding theme and location. The ontology and its instance data are deployed (with slower performance) when the user drills down into a query to discover more detailed information about a category or an instance (knowledge not stored in the gazetteer).

References

Alexandria Digital Library project, http://webclient.alexandria.ucsb.edu/.

Bowers, S. and B. Ludäscher, 2004, An Ontology-Driven Framework for Data Transformation in Scientific Workflows. In *International Workshop on Data Integration in the Life Sciences (DILS'04)*, March 25-26, 2004, Leipzig, Germany.

Fonseca, F., M. Egenhofer, P. Agouris, and G.Camara. 2002. "Using Ontologies for Integrated Geographic Information Systems", Transactions in GIS, 6(3), 2002.

Hill, L. L. 2000. Core elements of digital gazetteers: placenames, categories, and footprints. In J. Borbinha & T. Baker (Eds.), Research and Advanced Technology for Digital Libraries : Proceedings of the 4th European Conference, ECDL 2000 Lisbon, Portugal, September 18-20, 2000 (pp. 280-290). Berlin: Springer.

Hill, L. L. 2006. Georeferncing. The MIT Press. Cambridge, Massachusetts.

Jones, C., A. I. Abdelmoty, D. Finch, G. Fu, and S. Vaid. 2004. The SPIRIT Spatial Search Engine: Architecture, Ontologies and Spatial Indexing. In *Proceedings of the 3rd International Conference on Geographic Information Science*, pages 125–139.

Larson, Ray R. 1996. Geographic Information Retrieval and Spatial Browsing. In L. Cl, Smith and M. Gluck, eds., Geographic Information systems and Libraries: Patrons, Maps, and Spatial Information, 81-123. Urbana-Champaign: Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign.

Lutz, M. and E. Klien. 2006. Ontology-based Retrieval of Geographic Information, International Journal of Geographical Information Science.

Proposal to participate in the NCGIA specialist meeting on Digital Gazetteer Research and Practice

Krzysztof Janowicz
Institute for Geoinformatics, University of Muenster
Robert-Koch-Str. 26-28, 48149 Muenster, Germany
Webpage: http://ifgi.uni-muenster.de/~janowicz

Similarity-Based Identity Assumptions for Historical Places

The domain of cultural heritage is very heterogeneous; the themes or exhibits that museums and related institutions are concerned with range from history of science to various kinds of art and historical documents, and biodiversity. Accordingly, the number and type of preserved exhibits range from millions of collected organisms to a small number of valuable paintings. Creating and maintaining metadata about exhibits and historical facts in general gets increasingly important for scholars and curators in order to structure, manage, and query their own data. As long as metadata is used for internal workflows only (such as the preparation of an exhibition), each institution may develop and maintain their own schema and representation format; however, to refine and enrich their own knowledge base or to answer complex scientific questions, interchange with external sources becomes necessary. Cleaning up the local knowledge base is especially important because one needs to keep in mind that historical knowledge may be vague, incomplete, or even misleading. To support these tasks the Committee on Documentation (CIDOC) provides a well established and standardized core ontology (called CIDOC CRM; ISO 21127) [1] intended to annotate heterogeneous cultural heritage information to make it available in a machine-readable format (RDF) and reasonable way for knowledge integration, mediation and interchange. The long-term vision is publishing all annotated datasets through web services and therefore create a shared network of interlinked historical information enabling automatic metadata harvesting. The CIDOC conceptual reference model can be regarded as the underlying semantic level that provides meaning within the intended cultural heritage data infrastructure (which can bee seen analogously to a Spatial Data Infrastructure) by delivering a common metadata schema. Instead of trying to reach a community wide agreement on definitions for concrete entity classes (such as types of exhibits) the strength of CIDOC CRM lies in defining an abstract but interrelated vocabulary describing the fundaments of historical facts, namely established links (relations) between places, actors, objects, and events.

To make use of external data sources, however, a common language is not sufficient. It must be guaranteed that the collected metadata refers to the same real world phenomenon (which could be a historical place, person, event, or object) as the local datasets. Global authorities (such as the Alexandria Digital Library Gazetteer Server) provide unique identifiers and annotated datasets for some common kinds of real world phenomena. Scholars can refer to these global identifiers in addition to (or instead of) their local identifiers and therefore reduce maintenance effort and redundancy on the one hand and enable data interchange on the other. If compared datasets refer to the same global identifier and the scholar decides to trust the global authority as well as the external party that linked their dataset to the specific identifier, it can be assumed that the same real world phenomenon is meant.

Nevertheless, so far most datasets do not refer to global authorities and scholars must decide as the case arises whether the harvested information is relevant for their own knowledge base. There are several reasons for this:

- Knowledge about historical places is often vague and incomplete.
- Non-unique place names (even within the same area)
- Place names refer to cities, rivers, valleys, mountains, etc.
- Misinterpreted place names (e.g. 'Al Wahat' → oasis)
- Names also refer to varying geopolitical units (e.g. nomads) or prominent (artificial) landmarks (e.g. telegraph stations)
- Out-dated place or even country names (e.g. UDSSR)

Finally, the most significant reason why global identifiers provided by Gazetteers can only partially solve the problem of identity is that using Gazetteers to determine whether two datasets refer to the same real world place, presumes that all involved institutions have manually annotated millions of local datasets beforehand, which is not the case until now. Therefore an identity assumption assistant should support scholars in analysing the harvested metadata and returning promising datasets - in a way that the external datasets *probably* refer to the same real world place addressed by the local data. The identity assumption theory used by such an assistant should be non-rigid in a way that it returns a ranked list of estimations instead of trying to automatically conclude safe predictions from vague historical data.
If, in practice disambiguation via gazetteers and other global authorities (such as for historical figures) is often difficult, expensive and error-prone (especially for subordinate geopolitical units, events, actors, etc.) an identity assumption service should use the links established via the CIDOC CRM annotation between places, actors, objects, and events as additional *reference* points. In other words, taking Goodchild's geographic reality (geoinformation as a spatiotemporal location vector and an attribute/thematic vector [2]) and Kuhn's notion of semantic reference systems [3] into account, the underlying idea is to use thematic information as support for spatiotemporal reference. The same way as the spatiotemporal location vector is interpreted by a spatiotemporal reference system, thematic information is interpreted by a semantic reference system defined by CICOC CRM as a formal ontology, and similarity and classical (spatiotemporal & subsumption) reasoning as functions over this defined terminology. Proposing similarity as part of the puzzle of identity assumptions is drawing the metaphor from our geographical notion of location to the location within a network of historical facts and the spatial 'next-to' relation to a thematic one based on similarity assessments (see [4, 5] for further details on the identity assumption theory itself and similarity measurement between conceptualisations described in formal languages).

References

[1] Crofts, N., et al.: *Definition of the CIDOC Conceptual Reference Model (version 4.2)*. 2005.
[2] Goodchild, M.: Geographical data modeling. Computers and Geosciences (1992). 18(4) p. 401- 408.
[3] Kuhn, W.: Geospatial Semantics: Why, of What, and How Journal on Data Semantics III. Springer Verlag LNCS 3534 (2005) p.1-24
[4] Janowicz, K.: Towards a Similarity-Based Identity Assumption Service for Historical Places. In M. Raubal, H. Miller, A. Frank, and M. Goodchild, Eds. Geographic Information Science - Fourth International Conference. Springer Verlag LNCS 4197 (2006) p.199-216.
[5] Janowicz, K: Sim-DL: Towards a Semantic Similarity Measurement Theory for the Description Logic ALCNR in Geographic Information Retrieval. R. Meersman, Z. Tari, P. Herrero et al. (Eds.): SeBGIS 2006, OTM Workshops 2006, LNCS 4278 (2006) p. 1681 - 1692

Position Paper for **Digital Gazetteer Research and Practice Workshop**

Gazetteers and Geographical Information Retrieval

Chris Jones
School of Computer Science
Cardiff University, UK

Gazetteers are coming to play an increasingly important role in geographical information retrieval on the web. They enable users of transport timetables, routefinders, yellow pages, web mapping services and geographical web search engines to employ place names when specifying the geographical context of their requirements. The use of gazetteers in this role has also served to highlight some of their limitations with regard to the needs of the user. In practice, many queries that specify place names fail. One of the prime reasons for this is that the user may employ an informal or vernacular place name that is in common use but which is not recorded in the available gazetteers. In the UK, examples of such names are the "Midlands" the "Chilterns" and the "Wye Valley". The reason the name would not be recorded is that gazetteers tend to reflect an administrative view of the world with an emphasis upon places that have precise boundaries. Some gazetteers do record the names of topographic features such as mountains and valleys, but they are not usually accompanied by data that record an estimate of the spatial extent of the features. The existing gazetteers may also fail to recognise a name because they lack the required level of detail or geographical extent or simply because they are out of date.

There is a need therefore for richer gazetteers that reflect common knowledge of place names. Because of the high rate at which place names change or are introduced there is also a need to develop a system of interoperable web gazetteer services that reflect local and regional knowledge of places throughout the world.

For the purposes of geographical information retrieval it is possible to envisage an ideal situation in which there is a system of multilingual gazetteer services in which the content conforms to agreed methods for specifying: preferred and alternative names; the timeframe for use of names; an ontology of place categories; rich information on spatial context including geo-political and topographic hierarchies, coordinates in well defined reference systems, spatial relations to adjacent places, and spatial footprints at different levels of generalisation, with information on the nature of boundaries (precise / vague).

Vernacular names

The issue of incorporating vernacular names into gazetteers raises several challenges with regard to the source of the knowledge and methods for modelling and representing it. Where does knowledge of the names come from? There is a great deal of personal knowledge of place names that it is possible to envisage eliciting via some form of mass questionnaire conducted perhaps on the web. There is a considerable body of vernacular place name knowledge within textual documents. Many such documents are to be found on the web and web mining or web harvesting is therefore another route to knowledge acquisition.

Preliminary work on web mining for place name knowledge (by Purves, Clough and others) has demonstrated that simple web queries that include target vernacular place names result in the retrieval of web documents that refer to places that are co-located with the vernacular place, often being inside it. By performing a statistical analysis of the frequency of co-occurrence of place names with the target name it is possible to identify fuzzy regions of space that may approximate the extent of the place. Where the imprecise place contains few other places then it may be appropriate to attempt to identify other co-located topographic features (rivers, mountains, lakes, valleys, forests etc) that may be mentioned in association with the target place.

Several geometric modelling methods have been employed to represent the extent of imprecise places, in particular based upon the use of Voronoi diagrams, Delaunay triangulations (Arampatzis et al) and surface density functions. The latter surface modelling methods are notable for being able to represent the uncertainty of boundaries modelled by the frequency of occurrence of co-located places. Arbitrary precise boundaries can be generated from the surface models by choosing a threshold value of the surface, but it is a challenge to determine what value the appropriate threshold should take.

Building a system of gazetteers with extensive and rich geographical coverage can be expected to require a considerable degree of data integration in order to exploit the variety of data sources available. In addition to methods such as those referred to above, there are more conventional sources of place name knowledge within digital map products, their associated name lists or gazetteers and within geographical thesauri. These sources differ from each other with regard to the coordinate systems, accuracy and precision of geo-referencing, the form of the geometric footprint (points, polygons minimum bounding boxes for example), the nature of administrative hierarchies, the classification systems employed to describe the topographic type of the place, and language variation, along with inconsistencies in spelling and of naming the same places. At present when attempting to merge sources such as the Getty TGN with local gazetteer and topographic map-based names, major problems arise due to differences of the sort listed and there is a need to develop robust methods for integration. It may be useful in the short term to create some benchmark datasets containing expert-asserted equivalences between different represesentations of places in order to assist in evaluating automated methods for place matching and data integration.

References

Arampatzis, M. van Kreveld., I. Reinbacher, C.B. Jones, S. Vaid, P.D. Clough, H. Joho, and M. Sanderson, Web-based delineation of imprecise regions, *Computers, Environment and Urban Systems (CEUS)*, Volume 30(4), pp. 436-459.

Purves, R., Clough, P. and Joho, H. (2005), Identifying imprecise regions for geographic information retrieval using the web, *In Proceedings of GIS RESEARCH UK 13th Annual Conference*, Glasgow, UK, pp. 313-318.

Enhanced Gazetteer Development for Multilingual Geographic Information Retrieval of Natural Language Text

Ray Larson and Fredric Gey
University of California, Berkeley

Geographic information retrieval (GIR) from text is the subject of active research in the information retrieval research community. GIR focuses upon search and retrieval with a geographic component, e.g. *Find stories about cities near the Danube and Rhine rivers in Europe.* There have been GIR research workshops in 2004 (SIGIR, Sheffield UK), 2005 (CIKM Hanover, Germany) and 2006 (SIGIR Seattle USA). GeoCLEF (http://ir.shef.ac.uk/geoclef/) is a component track in the European Cross-Language Evaluation Forum (CLEF) which evaluates performance of Geographic Information Retrieval on multilingual text by creating test topics in multiple languages which are run against document collections in those languages For the GeoCLEF 2006 evaluation just concluded and results presented in Alicante Spain, the languages were English, German, Portuguese and Spanish (additionally, topics were translated into Japanese for cross-language search from that language). The document collections consisted of news stories from USA, Swiss and Germany, Portugal and Brazil, and Spain. The total number of documents being searched exceeds 1 million documents. GeoCLEF has emerges as the standard by which GIR for text research advances can be objectively evaluated.

Among the components of GIR search topics which differ from ordinary information retrieval are:

- Geographic challenge:
  <EN-title>**Cities within 100km of Frankfurt**</EN-title>
  <DE-title>**Städte im Umkreis von 100 km um Frankfurt**</DE-title>
  <PT-title>**Cidades a menos de 100 quilómetros de Francoforte**</PT-title>
  <ES-title>**Ciudades a menos de 100 kilómetros de Fráncfort**</ES-title>
  <JP-title>□□□□□□□□□ **100km** □□□□□□□</JP-title>
- Geographic location disambiguation for vaguely defined entities:
  <EN-title>**Scientific research in New England Universities**</EN-title>
- Geotemporal disambiguation for vague references
  <EN-title>**Credits to the former Eastern Bloc aka the Warsaw Pact**</EN-title>
- Approximate regional restriction:
  <EN-title>**Forest fires in Northern Portugal**</EN-title>

Research issues which have been explored by GeoCLEF participants and GIR researchers include named entity extraction in multiple languages, place name disambiguation, geographic hierarchy and expansion, as well as examining issues about the granularity of gazetteer information (e.g., when expanding queries using placenames derived from gazetteer lookup, should only major populated areas be used, or should all toponyms in

the referenced area). Researchers have also explored how to combine text ranking and geographic ranking schemes in retrieval.

Digital gazetteers form a critical infrastructure for GIR and for related Digital library applications. Obviously, for cross-language retrieval digital gazetteers must include place names in multiple languages, with appropriate point coordinates or footprints for the places. At UC Berkeley we have been involved in a variety of projects ranging from research and development of GIR ranking methods to the development of time and space-based retrieval systems for library catalogs and internet resources. Much of this work has been conducted in collaboration with the Electronic Cultural Atlas Initiative (ECAI). One of the goals of ECAI is to have an interactive Map based search and discovery front end so that the initial search for distributed metadata does not have to be a text search. This requires that there be geo-temporal metadata in the primary metadata for searching which will allow identification of the area of coverage of an object or collection and information about where to get the additional metadata and data required to provide the transactional geo-temporal browsing functionality. For ECAI the geo-temporal metadata serves as the union catalog for accessing a wide range of distributed transactional objects. Digital Gazetteers, in conjunction with the Time Period Directory developed as a prototype over the past 2 years, provide key elements for linking events, places and the people and subjects related to them. (See http://www.ecai.org/imls2004/imls4w/ The time period directory is a digital structure analogous to a digital gazetteer, but which associates named periods and events with date ranges, as well as referencing gazetteer information for the places associated with those time periods and events.) The combination of Time Period Directories and Gazetteers provides a metadata infrastructure for ECAI projects and systems.

Our interests in attending the meeting are twofold. First we will represent the needs of the cross-language IR research community (Ray Larson and Fred Gey are co-chairs of the GeoCLEF evaluation track), and second we will represent the interests of ECAI in gazetteer standards development and cultural and humanities applications of geo-spatial information resources. It is our contention that ordinary gazetteers are a necessary but insufficient component for resolving the problems of geographic search of natural language text and supporting the required access mechanisms for systems like those being developed for ECAI. Among the needed enhancements to Digital Gazetteers are:

- To enhance gazetteers with thesaurus-like co-references including the ordinary language expressions (Latin America, Middle East, etc.) which can be tied to exact gazetteer geographic components.
- Additionally historical names and footprints for places should be added whenever possible to provide a temporal dimension for gazetteers
- To link geographic information to its historical context, either by directly embedding it in the gazetteer, or by reference to time period directories.

Having attended (Larson) the previous gazetteer workshop in Washington DC, we look forward to an interesting and productive meeting.

Linda L Hill
University of California, Santa Barbara (emeritus)

Digital Gazetteer Research and Practice Workshop, December 7-9, 2006

In 1999, Mike Goodchild and I co-chaired a workshop on gazetteers—the Digital Gazetteer Information Exchange (DGIE) workshop, funded by the National Science Foundation. The workshop report includes a number of specific recommendations for future research and development activities. The highlighted conclusions were that there was an "immediate opportunity and requirement to coordinate the building of shareable digital gazetteer data in the interest of digital earth applications," that the temporal aspects of gazetteer data should get more attention, and that a gazetteer service protocol to support distributed gazetteer services should be developed.

In planning for this DGRP workshop, I see the developments that have taken place in the intervening seven years and the situation today against the backdrop of the DGIE report. Now there are more services based on gazetteer data, including interpreting geographic references in text (geoparsing), placename orientation of map views, placename access to online geospatial data, and communities developing shared gazetteers for their own purposes. There are more online gazetteer datasets, some with open contribution through web applications. The gazetteer data model developed by the Alexandria Digital Library (ADL) Project has been adapted for special gazetteer applications and a group at UC Berkeley has adapted it for representing named time periods, but no further work has been done (that I know of) to develop similar models for other geospatial and temporal data sets, such as meteorological events (e.g. hurricanes). The only networked example for gazetteers is the one based on the ADL Gazetteer Protocol giving access to both the ADL Gazetteer and to the ESRI Gazetteer. The OGC has published a gazetteer service protocol as a specialization of its Web Feature Service specification; the protocol is based on the model of gazetteer data formalized in the ISO TC 211 standard for *Geographic Information – Spatial Referencing by Geographic Identifiers*. There is a growing understanding of the role of gazetteers as components of information retrieval systems—gazetteers as a type of knowledge organization system (KOS)—but text-based information management and retrieval systems are still largely without a geospatially-enabled access mode and translation from placenames to coordinates. Some papers on gazetteers have been published, some conference presentations on gazetteer related research have been made, and with the publication of my book on *Georeferencing: The Geographic Associations of Information* by MIT Press there is now a substantial chapter on gazetteers available in book form.

Gazetteer research and practice draws from the expertise of diverse professions because placenames, classification schemes, and geospatial representations are all key components. The chief division is between text-based systems and geographic information systems. I think it is fair to say that currently those who work with text-based information resources and information retrieval systems are slowly integrating their placename georeferencing methods with geospatial representations and those who work primarily with geospatial information are realizing that their treatment of placenames and place classification would benefit from a more formal approach. Projects outside of the official toponymic authority agencies are realizing the complications and cost of building and maintaining quality

gazetteers. Such projects are very interested in using gazetteer data already collected and documented and developing ways to collect new place description information from knowledgeable sources.

The primary gazetteer research interests that I bring to this workshop are:

Components of Gazetteer Services:
1. An information retrieval test environment where footprint generalizations and similarity calculations can be tested for performance for given tasks, answering, for example, when bounding boxes are sufficient are for geospatial information retrieval.
2. Analysis of cross-walking options for feature type classifications, including automatic methods derived from the placenames themselves, for a 'gazetteer classification advisory service' that can be used to support gazetteer search interactions and gazetteer creation.
3. Development of software for gazetteer creation and maintenance based on community standards that can be customized for individual and organizational purposes.
4. Modeling of the temporal and spatial components of gazetteer data for applications in which the temporal aspects are on equal footing with the spatial.

Georeferencing as a Process:
1. Establishing college curricula and internships to educate GIS and LIS students in role of gazetteers in information services.
2. User studies for gazetteer services.

Interoperable Gazetteer Services:
1. Conflation of placename data from multiple sources for one place. This is a complex problem because all attributes of a place can vary: it may be known by different names, different terms may be used to represent its feature type, and different representations of its coordinate location may be used due to source, scale, and time period.
2. A test environment for gazetteer service interoperability, testing gazetteer service protocols and the suite of services needed to support discovery, search and retrieval. This would include a network of gazetteers accessible by a common gazetteer protocol and methods to obtain comparable performance data.

As a result of this workshop, I would like to see an organized and focused research effort to address gazetteer data and gazetteer service issues and specific efforts to teach the fundamentals of gazetteers and the services built on them through our professional educational systems.

Extending Gazetteer Service in Geographic Information Retrieval

A Position Paper

Naicong Li, Jordan Henk
The Redlands Institute, University of Redlands

Hill (2000) identifies three basic elements in a digital gazetteer: *name*, *footprint*, and *type* (or category). Gazetteer entry *types* are organized in a hierarchical classification system. The name and footprint properties of geographical features enable automatic association between non-spatial data (e.g. text documents) and the location of geographical features, whereas the feature type hierarchy provides automatic organization of the the spatially indexed data. Used in a geographic information retrieval system (GIR, Larson 1996), such spatial indexing makes it possible to answer thematic ("what") and locational ("where") queries. The "theme" in the query may be spatial, referring to a geographic feature (e.g. the City of Redlands) or to a category of such features (cities). It can also be non-spatial, referring to some property of the feature (e.g. the city's population). While the *name* and *footprint* properties of geographic features are formally defined in a gazetteer, the typical set of properties associated with a particular feature category often are not. Similarly, the typical set of relationships among geographic features (other than the *is-a* relationship, such as the *capital-of* relationship) are often not present in gazetteers. GIR systems could potentially benefit from explicit definitions of these properties and relations by using them to automatically index and later retrieve relevant data, especially in the case of highly structured data such as GIS datasets.

The Redlands Institute at the University of Redlands is conducting applied research in GIR using a prototype platform known as the Geospatially Referenced Information Portal (GRIP). GRIP was designed to facilitate knowledge synthesis across a disparate set of information resources (sources such as internet engines, local file systems, digital libraries, databases, etc. - all stored in heterogeneous formats, possibly with non-standardized structures). GRIP (which is being extended to support knowledge management tasks for a desert tortoise science and habitat management project) currently uses a project gazetteer with a feature type hierarchy modeled after the ADL Feature Type Thesaurus (FTT). This gazetteer has over 12,000 features extracted from recent versions of the USGS GNIS gazetteer (for the states of CA, NV, AZ, and UT); USGS Federal Land polygons augmented with boundaries for states, counties, state parks, etc.; and further supplemented with project-based GIS data. Using this gazetteer, GRIP supports basic spatial searches and basic queries (including theme and location).

Our users are often interested in the exact value or range of values of a certain property of a geographic feature (e.g. the actual tortoise population density of a critical habitat unit), or the statistics on these values, or retrieving a subset of geographic features of a certain type based on some specification of their property values. Sometimes the data value already exists in the project GIS dataset(s), sometimes GRIP needs to perform spatial and/or non-spatial processing to derive the answer. To handle such queries, we have adopted an approach similar to those described in Fonseca et al. 2002 and Lutz et al. 2006. We are prototyping using a light weight (not heavily axiomatized) project ontology for our user community. This ontology contains project-based concept categories (both spatial and non-spatial). For each concept category, we define a set of relevant properties (e.g. the 'area' of a 'habitat'), including definitions of the property value type and value range. We also define its relationship (of types *is-a*, *partitive*, *causal*, *functional*, *associative*, etc.) to objects of other categories when applicable. Data in the project database are linked to appropriate categories as instances of these categories. Such links are established by a process similar to "registration mapping" as described in Owers and Ludäscher 2004, and in Lutz and Klien 2006. The categories in this ontology thus can be used as an index into the data in the project geodatabase. The categories can also be used to index document metadata, and they may be used to annotate text documents (through recognition of their corresponding natural language expressions in the documents) to produce sub-document level metadata for various levels of discourse units such as word, phrase, sentence, paragraph, etc. This ontology is designed to import pre-established upper level ontologies including an ontology of spatial relations with definitions of topological relations, distance and direction. It also incorporates definitions of frequently used spatial and non-spatial operations which can be invoked when GRIP is processing the more complex spatial search

(Larson 1996, Jones et al. 2004). Using the properties, relations and rules defined in this ontology, GRIP will be able to deduce implied information through inferencing and reasoning.

GRIP exposes this ontology to its user in one of its query formulation interfaces. One advantage of ontology-driven query formulation is that the category, instance and property in the ontology may be used as a set of controlled vocabulary terms in query formulation (thus by-passing the need of disambiguating user queries in natural language). Another advantage is that this approach facilitates user exploration (and therefore discovery) of the information stored in the project database(s). A novice user, unfamiliar with desert tortoise habitats, may browse the ontology and learn about relevant factors (such as climate conditions, vegetation cover, soil, threats, etc.) through the property definitions for the concept of "Desert Tortoise Habitat". The user may also obtain a more comprehensive knowledge of threat factors by browsing all the sub categories of "Threat".

Our research is intended to produce a generic platform in which a project ontology and a project gazetteer may be integrated to enhance geographic information retrieval. Because of its relatively simple structure and the use of an efficient indexing scheme, the gazetteer has a much faster performance time during a search. The gazetteer is also sufficient for answering the most common types of queries regarding theme and location. The ontology and its instance data are deployed (with slower performance) when the user drills down into a query to discover more detailed information about a category or an instance (knowledge not stored in the gazetteer).

References

Alexandria Digital Library project, http://webclient.alexandria.ucsb.edu/.

Bowers, S. and B. Ludäscher, 2004, An Ontology-Driven Framework for Data Transformation in Scientific Workflows. In *International Workshop on Data Integration in the Life Sciences (DILS'04)*, March 25-26, 2004, Leipzig, Germany.

Fonseca, F., M. Egenhofer, P. Agouris, and G.Camara. 2002. "Using Ontologies for Integrated Geographic Information Systems", Transactions in GIS, 6(3), 2002.

Hill, L. L. 2000. Core elements of digital gazetteers: placenames, categories, and footprints. In J. Borbinha & T. Baker (Eds.), Research and Advanced Technology for Digital Libraries : Proceedings of the 4th European Conference, ECDL 2000 Lisbon, Portugal, September 18-20, 2000 (pp. 280-290). Berlin: Springer.

Hill, L. L. 2006. Georeferncing. The MIT Press. Cambridge, Massachusetts.

Jones, C., A. I. Abdelmoty, D. Finch, G. Fu, and S. Vaid. 2004. The SPIRIT Spatial Search Engine: Architecture, Ontologies and Spatial Indexing. In *Proceedings of the 3rd International Conference on Geographic Information Science*, pages 125–139.

Larson, Ray R. 1996. Geographic Information Retrieval and Spatial Browsing. In L. Cl, Smith and M. Gluck, eds., Geographic Information systems and Libraries: Patrons, Maps, and Spatial Information, 81-123. Urbana-Champaign: Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign.

Lutz, M. and E. Klien. 2006. Ontology-based Retrieval of Geographic Information, International Journal of Geographical Information Science.

A Framework For Inferring Spatial Locations And Relationships From Text

Inderjeet Mani, Dave Anderson, and Janet Hitzeman

Systems that interpret spatial information in natural language text need to deal not only with 'absolute' references (e.g., "*Rome*", "*Rochester, NY*"), but also relative references ("*thirty miles north of Boston"*, "*an underpass beneath Pushkin Square*"). Current approaches to extracting information from text have made excellent progress using a methodology of first developing an annotation scheme for marking up expressions of interest with various features, and then training machine learning algorithms to reproduce the annotation. Earlier research along these lines has yielded success in resolving absolute and vague time expressions in different languages using the TIMEX2, annotation scheme (timex2.mitre.org, Mani et al. 2005), and temporally situating text mentions of events using the TimeML annotation scheme (www.timeml.org, Mani et al. 2006). We have recently begun a 3-year project to apply such a methodology, for the first time, to the automatic interpretation of spatial expressions in natural language texts in English and Chinese. Here we describe aspects of our project, building on our work to date, that are relevant to the themes of the workshop.

SpatialML Markup Language

We are currently developing a markup language for spatial expressions called SpatialML that provides a semantically-based scheme for marking up spatial expressions. It is being applied to a variety of different types of texts (including news, weather forecasts, route descriptions, geographical descriptions, etc.) , with a corpus with this markup (currently already marked up with place names disambiguated with gazetteer-related features) being distributed and used as training data by various machine learning algorithms. For example, in the case of "*an underpass beneath Pushkin Square*", "*underpass*" would be tagged in SpatialML as a feature of a particular type based on an existing feature ontology, "*beneath*" would be tagged as a *signal* with a value for a *topological relation* feature, and "*Pushkin Square*" as a particular *place* with a value for a *geo-coordinate* feature.

Place Name Disambiguation

A common way of referring to space is of course in terms of proper names. Accurate disambiguation of place names in text in terms of points and regions on a map are dependent on gazetteers with geo-coordinates and geographic inclusion information. Large gazetteers increase the degree of ambiguity; for example, there are 1420 matches for the name "*La Esperanza*", according to the GeoNames Database from the National Geospatial-Intelligence Agency (NGA). A recent study (Garbin and Mani 2005) on 6.5 million words of news text found that two-thirds of the place name mentions that were ambiguous in the U.S. Geological Survey's GNIS gazetteer were 'bare' place names that lacked any disambiguating information in the containing text sentence.

Information extraction systems can use disambiguation rules based on human intuition as well as rules discovered by programs trained from disambiguated examples. Since it is expensive to generate adequate samples of training data, research has tried to trade off

quality of training data against quantity. The Garbin and Mani study showed that nearly four out of five place names were accurately disambiguated when a machine learning program was trained on 11.7 million words of English news text that had been automatically disambiguated using hand-coded heuristics. The success of such heuristics is dependent in part on the gazetteer used; a larger gazetteer will lead to more ambiguity. MetaCarta (www.metacarta.org), one of the well-known systems for place name tagging and disambiguation, for example, has a gazetteer of roughly 10 million entries.

Gazetteer Integration

Gazetteers are fundamental to geographical information extraction. Earlier work has explored harvesting and semi-automatic integration of multiple gazetteers to support place-name identification and disambiguation from text. We have experimented with spatial distance and geographical feature-based entries in matching entries across gazetteers. In related work, we have developed algorithms for matching transliterated variants of person names based on both sound and spelling, which are used to generate training data for machine learning approaches that learn costs of string edit operations. We will explore the impact of such name comparison approaches in terms of merging gazetteers as well as gazetteer lookup.

Spatial and Temporal Reasoning

Today's information extraction systems today reason very little, if at all, about space and time, e.g., systems cannot represent the fact that the same entity cannot be in two places at the same time. This results in extracted data that is highly incomplete, requiring considerable interpolation and extrapolation by a user. For the problem of temporally situating events, we have discovered that temporal reasoning, in the form of transitive closure over annotated qualitative temporal relations (precedence, inclusion, etc.), can be used to dramatically expand the amount of training data (Mani et al. 2006), and we expect similar benefit from spatial closure, e.g., over relations such as connection and inclusion of spatial regions. Information from domain databases will also be used to constrain information extraction results, so that in the case of "*an underpass beneath Pushkin Square*", the potential candidate underpasses can be identified and displayed on a map.

References

Inderjeet Mani, Marc Verhagen, Ben Wellner and James Pustejovsky. (2006). Machine Learning of Temporal Relations. Proceedings of the Association for Computational Linguistics (ACL'2006), Sydney, Australia.

Eric Garbin and Inderjeet Mani. (2005). Disambiguating Toponyms in News. In Proceedings of the Human Language Technology Conference (HLT-EMNLP'05).

Inderjeet Mani, James Pustejovsky, and Rob Gaizauskas. (2005). The Language of Time: A Reader. Oxford University Press.

How Gazetteers Work:
Cultural and Linguistic Influences on the Georeferencing Process

David M. Mark
NCGIA & Geography
University at Buffalo, Buffalo NY, USA

As noted in the workshop introduction, georeferencing by naming is universal, and has existed for a long time. Knowing the names has long been a traditional way of demonstrating ownership of country. Before written language and graphic maps, geographic information was often preserved and transmitted through stories, which often also included other important cultural information or moral codes. Place names, and the places themselves, formed ‘retrieval keys’ for the associated information. Keith Basso's book, "Wisdom Sits in Places" explores such a system still in use by the Western Apache, and there are many examples from Homer's Iliad and Odyssey and the Viking Sagas to the Navajo origin stories and traditions in many and cultures.

Today, we continue to use placenames in our everyday lives, to communicate about locations. But names also play a critical role in accessing information via the Web. Gazetteers provide a link between place names and locations externally, first in printed forms and more recently using digital computers.

In this paper, I will present some thoughts on the cognitive process of how place names *work*, and how they connect names to locations. I will then outline various places in this process where cultural, linguistic, and national differences influence these processes and provide challenges for multilingual gazetteers.

A gazetteer consists of associations among placenames, categories, and footprints (Hill et al., 1999; Hill, 2000). But each of these components is itself complex. The semiotic triangle (Ogden and Richards 1923) is used here as a key model to provide insight into how meaning and reference work in language in general and in gazetteers in particular. Normally, the semiotic triangle shows how a sign has meaning. The sign or symbol, perhaps a word or a name, is linked by a concept (mental structure) to an instance or category in the world. The link between sign and referent is dashed to illustrate that normally, words are linked to things in the world only through concepts.

For the placename corner of the gazetteer triangle, the operation of the triangle is simple. The proper name, "The Matterhorn" or "Santa Barbara", is the sign. The referent is a particular feature in the environment or landscape. Our concept of the appearance and characteristics of The Matterhorn or Santa Barbara allows a two-way associative link between those.

The category corner of the gazetteer triangle has a similar structure. The sign is a landscape generic or feature code such as 'mountain' or 'city'. The concept is the general idea of a mountain or city in general. And the referent is a set of all mountains or cities or a representative sample.

The footprint corner of the gazetteer triangle is the only one that does not seem to expand into a semiotic triangle. Rather, it needs an idea of how entities of that type are bounded (crisp, graded, what gradient; cf. Montello *et al*., 2005), and the application of the appropriate boundary concept

to the particular entity being referenced.

Cultural and Linguistic Differences in Placenames, Categories, and (Perhaps) Footprints

It is clear that many of the above components vary by language and by culture. Most of us are familiar with different language-specific placenames. Firenze becomes Florence, London becomes Londres, and Cologne becomes Köln or Colonia; even though these are usually just symbol substitutions and refer to "the same thing", that will not always be the case. That is because categories also vary by culture and language, sometimes by splitting or merging (for example, English 'river' splits (more or less) into 'fleuve' and 'rivière' in French) but often in more complicated ways (Mark, 1993; Mark and Turk 2003; Mark *et al*., in press). Although geographic entity delimitation has not been studied cross-linguistically, it seems quite possible that the footprint of "the same" feature might be significantly different in different cultural and linguistic contexts, especially for features with graded boundaries such as mountains (Smith and Mark 2003).

The presentation will some examples and suggest some approaches to solving the multilingual and multicultural aspects of gazetteers.

References

Hill, L., 2000. Core elements of digital gazetteers: placenames, categories, and footprints. Proceedings of the 4th European Conference on Research and Digital Libraries. Lecture Notes in Computer Science 1923, pp. 280.

Hill, L., Frew, J., and Zheng, Q., 1999. Geographic Names: The Implementation of a Gazetteer in a Georeferenced Digital Library. Digital Library 5(1), Corporation for National Research Initiatives, Technical Report: january99-hill.

Mark, D. M., 1993. Toward a Theoretical Framework for Geographic Entity Types. In Frank, A. U., and Campari, I, editors, *Spatial Information Theory: A Theoretical Basis for GIS*, Berlin: Springer-Verlag, Lecture Notes in Computer Sciences No. 716, pp. 270-283.

Mark, D. M., and Turk, A. G., 2003. Landscape Categories in Yindjibarndi: Ontology, Environment, and Language. In Kuhn, W., Worboys, M., and Timpf, S., Editors, *Spatial Information Theory: Foundations of Geographic Information Science*, Berlin: Springer-Verlag, Lecture Notes in Computer Science No. 2825, pp. 31-49.

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.

Ogden, C. K., and Richards, I. A. 1923. *The Meaning of Meaning*, Harcourt, Brace, and World, New York.

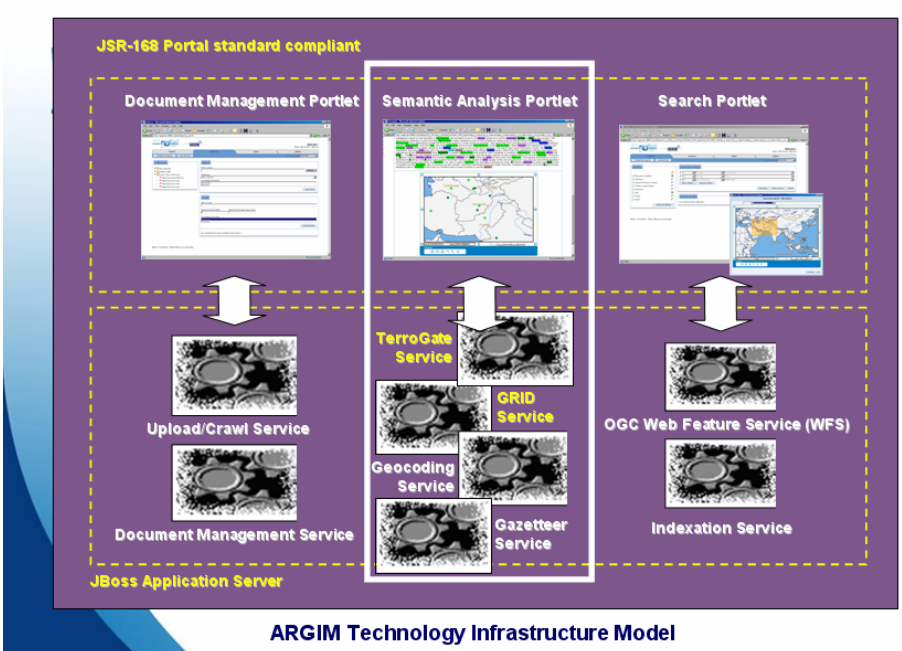
Smith, B., and Mark, D. M., 2003. Do mountains exist? Towards an ontology of landforms. *Environment and Planning B: Planning and Design*, 30(3), 411-427.

Marc-André Morin
System Architect
Defence R&D Canada

Expression of Interest in the topics of the Meeting

Defence R&D Canada (DRDC) is an agency of the Canadian Department of National Defence. Its mission is to improve Canada's defence capabilities, through research and development. My R&D area of expertise is the Knowledge and Information Management, specifically Geographic Information Systems (GIS). I am involved in a major applied research project: ARGIM, a GEOINT solution integrating knowledge exploitation and geospatial technologies.

ARGIM stands for Applied Research for Geospatial Information Management. The Project Leader is Dr Alain Auger, a Defence Scientist with expertise in computational linguistics. The ARGIM project leverages on both natural language processing and GIS expertise in order to develop a new GEOINT framework, or a wide toolbox, based on free and open source software to rapidly integrate, deploy and evaluate new Information & Knowledge Management concepts and technologies.



ARGIM Technology Infrastructure Model

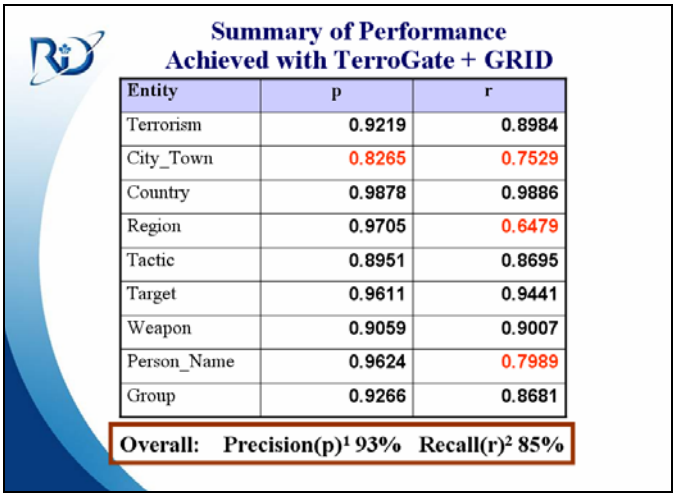
Designed and implemented by our team, the current prototype relies on Service Oriented Architecture (SOA). The portal includes:

1. Document Management Service for grabbing, structuring and sharing sources of information;
2. *Natural Language Processing (NLP)* services for semantic-based text search and analysis, including automatic annotation of geographically-related entities in unstructured documents;
3. GIS capabilities for indexation, retrieval, visualization and analysis of spatio-temporal information. Notice that specifications from the Open Geospatial Consortium (OGC) are taken into considerations.

NLP services include:

1. TerroGate service, a new information retrieval system dedicated to the terrorism domain recognition (tactics, weapons, targets, groups, etc.);
2. GRID service, a geoparser for geographic pattern-based recognition;
3. Geocoder for assigning geometries to representative geographic location terms contained in texts;
4. Finally, gazetteers for feeding all services described previously in their specific activities.

TerroGate and GRID have both been developed by DRDC Valcartier. They are both named entities extraction technology based on the free and open source software called GATE (http://gate.ac.uk/), a human language processing system to develop linguistic-based technologies. In fact, the generic "Named Entities Extractor" module of GATE has been modified, specialized and trained in order to increase its performance over electronic texts related to the terrorism and the geospatial domains.



Summary of Performance Achieved with TerroGate + GRID

Entity	p	r
Terrorism	0.9219	0.8984
City_Town	0.8265	0.7529
Country	0.9878	0.9886
Region	0.9705	0.6479
Tactic	0.8951	0.8695
Target	0.9611	0.9441
Weapon	0.9059	0.9007
Person_Name	0.9624	0.7989
Group	0.9266	0.8681

Overall: Precision(p)[1] 93% Recall(r)[2] 85%

As shown in the figure above, poorest results in terms of precision[1] and recall[2] are mostly related to geographic named entities. Precision for extraction of cities and towns within text documents is relatively low essentially because of several partial matches. For example, "Washington" could be tagged where it should have been "Washington D.C." Therefore, precision could be increased by improving pattern matching rules, grammar and algorithms. However, to increase recall relative to cities, towns and regions – where a lower percentage is linked to many locations or place names that are missing in our gazetteer – we have to focus on richer gazetteers and domain ontologies and that will help us to complete gaps within our in-house gazetteer. To achieve that goal, the Alexandria Project is probably the best source of information. Thus, the experimentation of the ADL Feature Type Thesaurus and the ADL Gazetteer are probably a good start to improve and redesign our geospatial knowledge domain.

Proper identification and georeferencing of information with a geographical context is a big challenge, and digital gazetteers can be part of the solution, and maybe, part of the problem… Loading the whole gazetteer in memory to conduct natural language analysis is unfeasible and useless. To do so, we must categorize location-related pattern matching rules and gazetteer's features according to a scale of visibility. For instance, it is totally inadequate to process a grammar rule responsible for detecting zip/postal codes and load in memory all street names related to a city when a document is related to world wide topics, such as earthquakes. Thus, our geoparsing application needs to support and integrate more than one gazetteer, where each one will have its own specificity. It is obvious that digital gazetteers play a key role within our infrastructure, not only for georeferencing place names but also to feed our parsing applications like TerroGate and GRID.

Finally, geosemantic ambiguities are a real dilemma for georeferencing or geocoding place names. If we do a search for "Paris" by using the ADL Gazetteer Client, we get 57 matches. How can the machine choose the right one? Also, how can we georeference assertions such as: "20 miles north of Santa Barbara"? This is what ARGIM and also The Alexandria Textual Geospatial Integration project (TGI) try to address. It will be very interesting to put this subject on the table for an in-depth discussion in order to identify opportunities for collaborative efforts.

[1] **Precision** ratio is the proportion relevant named entities among the retrieved ones
[2] **Recall** ratio is the proportion of retrieved named entities among all relevant ones

Ruth Mostern
University of California, Merced
November 17, 2006

A Historian's Perspective on Gazetteers and Interoperability

Position Paper for Submission to the International Workshop on Digital Gazetteer Research and Practice

If a historian wishes to study the political history of the region south of the Caspian Sea, she might begin with the Achmenaeid Persians, architects of the world's first great empire; trace the history of the region through its dominion by the Greek speaking Selucid and Parthian empires; chart its return to Persian rule during the Sasanian Empire, and note its conquest by the Arab Umayyad Caliphate and their Buwayhid successors. In medieval times, she would observe that the region was occupied by the Turkish Khanate of Khwarizm, the Mongol Il-Khanate, and the Turko-Mongol Timurids before reverting to Persian rule under the Safavids in the sixteenth century. With the emergence of nation-states in the eighteenth century, she would note that the region south of the Caspian became known as Persia, a name it maintained until the founding of Iran in 1921. As imperial fortunes shifted, the boundaries around and within the region south of the Caspian Sea changed dramatically. And, as suzerainty over the region changed hands between overlords of many linguistic backgrounds, place names changed as well. Often, places in this region were known simultaneously by names in many languages at the same time; naming systems that could represent different and conflicting perspectives about power and dominion. The historical trajectory of Persian political geography is hard to map with detail and precision, but is reasonably easy to model in a gazetteer.

Considered simply in their basic function as indexes of named places, gazetteers are an essential tool of historical geography, and, by extension, a crucial element of any kind of historical research whatsoever. Much geographically focused historical research, particularly on eras prior to the nineteenth century, is based on sources that are saturated with place names, but ones that can be assigned geographical coordinates only with difficulty and a very high degree of ambiguity. By contrast with historical GIS *per se*, with its premium on place, gazetteers have the capacity to liberate historical geographers from a preoccupation with location. They allow scholars to save untold quantities of time, money and frustration attempting to precisely georeference historical places and their indeterminate boundaries, while allowing us to accomplish what we do best: investigating how place and space made meaning to people. In a way that would be very difficult for a historical GIS system, a gazetteer can:

- model how places were related to one another (as components of a hierarchy or as nodes in a network, for instance)
- associate multiple coexisting names with one another through time and in many languages
- structure information about the changes to names, locations, feature types and relationships for any place
- link all of this information to sources, in the tradition of rich attribution expected by scholars of history and culture.

Why, at that point, should historians and humanists be encouraged to create gazetteers at all, rather than writing monographs, as our disciplinary tradition and colleagues

encourage? There are several answers to that question. First, as Willard McCarty has recently argued for the digital humanities in general, the creation of databases and taxonomies has the effect of "rendering knowledge problematic." McCarty suggests that "the twin computational requirements of complete explicitness and absolute consistency" opens up a space for the scholar "to refine an inevitable mismatch between a representation and reality" in a way that other modes of production do not.

Second, a gazetteer is an extraordinary reference work and research tool for historians. A search of the Alexandria Digital Library for my hometown of Merced, for instance, produces a list of 80 results in the United States, Latin America and the Philippines—a rough map of the Spanish cultural world. Once the resource is complete, a search of the China Historical GIS for Beijing will yield a list and map of the dozen or more places that have been designated as a northern capital at any time in Chinese history.

Finally—and turning explicitly to the topic of this workshop session—historical geographical information structured into a digital gazetteer has the potential to have an immense impact if it is incorporated into an interoperable gazetteer system. A book on a shelf or a free-standing database on line about Sasanian or imperial Chinese geography would be interesting only to specialists. However, a global gazetteer that includes this data would be useful to people asking a wide range of questions.

An interoperable gazetteer system that incorporates historical information has one critical characteristic. That is, it will necessarily consist of many small component resources that have been created by hand, by individuals or small groups of specialists. A specialist on Tibetan Buddhist temples may be able to create an extraordinary gazetteer within his domain by trolling historical sources and archaeological reports, but would have no way to do the same for any area outside of his linguistic and scholarly expertise. Systems and services that incorporate historical gazetteers need to develop tools that can allow small gazetteers to be searched and used together, including disambiguation protocols and feature type cross-walks. Gazetteer developers need to be educated about lowest common denominator content standards and service protocols that will allow their efforts to be visible to such a system.

Finally, while gazetteers are super-powerful reference tools in their own right, they will find their greatest uptake as components of information systems. In the domain of contemporary place name information, we are aware of Google Maps and many other ubiquitous services. In the historical domain, gazetteer services should be a standard component of the resources that I think of as Atlases. Gazetteers can be incorporated with digital library services such as catalogue searching and links to images, timelines, texts or historical censuses. All developers in the digital humanities should be attuned to the presence and value of digital gazetteers, and design spatially aware systems that can incorporate place name search and map display.

Subject Discussion Areas of Interest

A. Components of gazetteer services

Through our GeoCrossWalk project and associated GeoParser development we have specific interests in:

- database population and creation issues related to multiple scales and multiple data sources
- feature typing (esp in multilingual contexts)
- accommodation for the variations and repetitions of placenames on a worldwide basis, including exonyms
- treatment of temporal issues

B. Georeferencing as a process

EDINA has been endeavoring to build georeferencing into the process of resource management within the UK academic network infrastructure by providing resources and tools to assist in explicit georeferencing. These include:

- geoparsing technologies
- middleware services and servers
- spatial data discovery services including registries and catalogues

C. Interoperable gazetteer services

EDINA builds services and projects on open standards and interoperability principles. We are active Open Geospatial Consortia (OGC) members and are involved in a range of data harmonization and interoperability initiatives, including:

- SOA approaches
- distributed federations of authority name files
- authentication and authorization issues including geoDRM

EDINA also has an interest in Grid computing (eScience) and emerging relationships with OGC and other geospatial interoperability standards.

Expression of Interest in Workshop on Digital Gazetteer Research and Practice

Susan Stone, University of California, Berkeley

My interest in gazetteers is a result of my longtime interest in two different areas of information science and digital libraries research. Since studying for my master's degree in Information Science at Berkeley in 1988 and throughout my work at Berkeley in the Information Server Project in the early 1990s and later the Museum Informatics Project, I have been interested in information system interoperability. My second interest is in language issues and multilingual data. This interest stems from my bachelor's degree in Chinese and my first master's degree in an interdisciplinary program involving sociolinguistics and education, where I specialized in writing systems and scripts.

In the context of the gazetteer workshop, I am most interested in issues involving multilingual and international gazetteer content. In the area of interoperability, I am interested in distributed interoperating gazetteers, gazetteer search and retrieval protocols, and integration of gazetteer data from multiple sources. In my professional role, I am interested in incorporating gazetteer lookup into Berkeley campus collection databases and other information systems. I am also interested in looking at gazetteer development in the context of new technologies and architectures for collaboration.

I first learned about spatial indexing in about 1992, when Linda Hill spoke at a conference of the American Society for Information Science and published a paper in the conference volume, which I edited with Michael Buckland. When I began working the Museum Informatics Project at Berkeley soon after, we used thesauri, such as the Getty Thesaurus of Geographic Names, for standardizing locations in databases of museum collections. We also began to collaborate with the campus GIS center with an eye toward mapping specimen data and other data in campus collections. It was as a participant, representing the Museum Informatics Project, in ECAI, the Electronic Cultural Atlas Initiative, where I actually met Linda Hill and became involved with gazetteers. I participated in ECAI's NSF-funded gazetteer project, which sought to extend the ADL standard to accommodate historical, cultural, and multilingual data. Later, I was part of a research group at Berkeley with Ruth Mostern (then of ECAI, now of UC Merced) and a graduate School of Information Management and Systems student, Melanie Feinberg, to model time period data to complement location data in gazetteers.

In September 2001 through August 2002, NSF funded the project titled "A Multilingual Gazetteer System for Integrating Spatial and Cultural Resources," in which ECAI worked with Academia Sinica of Taiwan to extend existing gazetteers standards to accommodate the cultural and historical resources that ECAI and its collaborators have been developing. For our prototype, we worked on a gazetteer of Taiwan. We worked with location data collected by Academia Sinica field workers, along with the data for Taiwan in the ADL gazetteer and Taiwan data from the Getty Thesaurus of Geographic Names. Data were in Chinese and English, and I developed a testbed in which we were able to search and deliver records from the combined data using the ADL protocol and code from UC Santa Barbara. Part of our work in this project involved investigating a suitable feature type thesaurus for the types of locations researchers in the humanities were most interested in, one that could

be translated into multiple languages (initially, Chinese). Although we did not come to a consensus on how to create a thesaurus that would translate readily into multiple languages, given the different semantic ranges of terms in different languages, one suggestion was to develop a multilingual thesaurus of concepts that could be used to describe locations at a very general level, while also using more specialized thesauri for classification that might not be available or easily translated into multiple languages. Since work on cross-language searching and semantic ontologies has progressed since our NSF project, I believe the issue of multilingual feature types should be revisited in the context of distributed, interoperable gazetteers.

A special focus of the ECAI project was the instability of topoynms over time and the interdependence of time and space in gazetteers. In that regard ECAI came up with suggestions for the ADL gazetteer protocol that involved incorporating time elements into all components of a gazetteer entry, adding more source fields for data acquired from multiple sources, and adding fields for uncertainty—regarding names, feature types, as well as locations. The uncertainty fields that were suggested for researchers working with incomplete descriptions in historical texts proved to be of use for the biological researchers translating location descriptions into spatial coordinates, such as the team now working on the BioGeomancer project, which seeks to create tools for georeferencing and mapping biodiversity data.

Over the past year or so I been an observer of the work of the BioGeomancer project and have been in contact with some members of the team on language issues. When that project is completed, I would like to see the work they have been doing and the software they develop applied more generally to the groups in the humanities and the collections that I work with in Berkeley's Museum Informatics Project.

In sum, I have been interested in working with gazetteers for a long time and I would like to participate in an effort to create a new research agenda and to put into practice at the University of California the results of current and ongoing research and development efforts.

Expression of Interest — Björn Svensson

Name: Bjorn Svensson
Current Position: Project Manager
Company: Environmental Research Institute (ESRI).
Years with current company. 6

In 2002 I collaborated with Linda Hill and Greg Janée on the design of the ADL Gazetteer Protocol, but my professional interest in gazetteer and placenames probably started in 1997 when working on the South African National Spatial Infrastructure Framework. As part of setting up the first African FGDC clearinghouse the seemingly simple concept of geographic metadata and place and theme keywords turned out to be an intriguing matter that still fascinates me.

I work as a project manager at ESRI's ArcWeb Services group where, among other things, I'm in charge of the "ESRI World Gazetteer" which is used in different ESRI products. The ESRI World Gazetteer is one of several gazetteer and gazetteer-style databases that are provided through ArcWeb Services. It is used both by desktop GIS software like ArcMap and ArcGIS Explorer, as well as web browser applications, for example MapMachine by National Geographic and Where's Yours by Nature Valley. The ESRI World Gazetteer is available for free for certain non-commercial usage.

Currently we use our gazetteers primarily as a means to find a location (preferably with a footprint) and to quickly zoom in to the location in question. This could be either in web applications or desktop software. In this context, a "location" is primarily used to be able to zoom in a map to a certain area (using the location footprint), or to find certain data about a specific placename. This location finding functionality can be divided into two main categories:

1. Traditional gazetteers – type in a placename
    - returns Placename, with Type, and Spatial Footprint
        - ESRI World Gazetteer
        - Special Gazetteers for U.S. landmarks, World Postal Codes etc. Note that the main reason to keep these as separate gazetteers is for royalty purposes.
2. Addresses – supports not just a "street address" but also an IP address, a domain name or a "phone number" – all of which can be thought of as placenames/locations.
    - Returns "placename" with point location and other attributes where available (but no bounding box)
        - ArcWeb currently support street address geocoding for Australia, New Zealand, many European countries, United States and Canada.
        - Phone Numbers for the United States
        - IP addresses

What we're interested in pursuing is in general more focused on practice than theory/research. For example, here are a few very interesting topics (in no special order):

1. Implementing a "standard", or at least "interoperable", thesaurus instead of local thesauri.
2. Better implement alternative names, making it clear when it is the identical feature with different names.
3. Better implement and support multiple languages (and scripts). Both for queries in a specific language, but also for searching and presenting names in other languages.
4. Further integration between gazetteer feature and GIS feature (with exact boundary). This would also help in updating/maintaining the gazetteer.
5. Further and smarter integration between gazetteer feature and different hierarchies (for example Goleta, Santa Barbara County, State of California, United States).
6. Further integration between gazetteer feature and attributes from both non-GIS and GIS data (for example, 2006 median income for Alaska).
7. Adding historical / time dimension to data.
8. Investigating and potentially using "standard" protocols in our applications, for example the new OGC Gazetteer Protocol or the ISO9112 Standard.
9. Improved ranking of search results, taking into account additional information from the user situation.
10. Reverse gazetteers – based on point or polygon tell me the name of a place, taking into account context to return the appropriate level of detail and type.
11. Improved solution for creating, and using, your own gazetteer. A user could create their own gazetteers either from spatial or non-spatial databases, and then being able to host and serve them on a managed service platform and served out as a standard gazetteer.

I hope that this workshop with its presentation, discussions, and interactions will further our efforts in these directions.

MAP LINK

Map Link The Comprehensive Source of Maps & Geographic Information

26 October 2006

Michael Goodchild
UCSB
Santa Barbara, CA

Dear Dr. Goodchild,

Thank you for squeezing me in to the Digital Gazetteer workshop at such short notice.

As you know, Map Link is unique. We maintain an inventory of thousands of map and atlases from every corner of the globe. In the process of buying and re-selling this material Map Link must collect and manage metadata for every item. This means every sheet in a set of topographic maps. We collect basic data such as: scale, sheet size, format, date of edition, sheet name, a simple LC classification; as well as data relating to acquisition, cost and sales information, and a thumb-nail image of each map for display purposes. We have metadata for over 90,000 different maps and atlases.

Each of the three workshop components is an area Map Link has had at least some experience that may prove interesting, if not valuable to the discussion.

“Interoperable gazetteer services”

Map Link has tested a variety of database management tools to effectively store and manage map metadata internally. In the last few years our Internet presence has come to offer the full range of our holdings. As you might imagine, this presents us with many challenges, so our approach was to limit direct sales to resellers. These resellers presumably knew how to lookup the products they needed. Yet, I think we always knew the key to sales is in replicating the inquiry of a good map dealer, or librarian, in getting to the needed item quickly and efficiently. Our commercial demands occasionally mirror the conditions other managers of spatial data may face. Indeed, I believe we have a lot to learn from each other, particularly in the way users approach the data. Gazetteers, properly designed and deployed, are one key tool for efficient data lookup.

“Components of gazetteer services”

We have experimented with some advanced indexing and metadata collection techniques. When we first began to collect data for the Internet it was for a Geosystems (now Mapquest.com) retail web store in 1997. In addition to the foregoing metadata, we also started recorded lat/long coordinate values for each sheet. As many of the maps we sell use an unmanageable or detectable grid system, we also began to set down a list of important place names for these sheets. This rapidly became an untenable prospect, and we backed off to simply record the already mentioned metadata.

“Georeferencing as a process”

Today, Map Link maintains an inventory database with very simple and limited query tools. Now that our new, direct-to-the-consumer web site is becoming fully populated, this consumer (data user, etc) expects more intelligence and more natural search tools. We expect to be using digital gazetteers to aid in product lookup. We have already begun to recognize many issues that may be of interest to your group. We will be listening for new ideas!

My background has always been in the commercial, or private sector of the map business, but I have worked closely with mapping agencies, standards boards, and spatial data committees. Currently I manage publishing, data acquisition, and national mapping product acquisition at Map Link. Since our business is in buying and re-selling hard-copy maps from publishers all around the world, this puts us in contact with every map publisher and map-issuing agency. In addition, since we sell hard-copy topo mapping for the entire planet, we acquire this material directly from each national mapping agency. As we continue to manage our contacts with each of these, they may prove useful in obtaining place name information, and metadata to further the digital gazetteer initiative.

My long-term work with geographic names authorities and standards committees is a function of representing the map industry and their interests. Most map publishers sincerely want to provide the user with the best information possible, but lack the knowledge and direction. A presentation by a representative from Google Earth at a recent Council of Geographic Names Authorities meeting in Boulder illustrated this fact. The results of your workshop will be reported back to map publishers and other spatial data providers around the world.

The Council of Geographic Names Authorities, The United Nations Group of Experts in Geographic Names, The International Cartographic Association, and other groups have welcomed and encouraged my participation so that I may report back to the users of this material--namely, the publishers and users. Having worked with these groups, I understand the availability and many of the design features of data sets around the world.

Thank you very much for this opportunity to participate in this important workshop.

Regards,
Will Tefft
Map Link

Background to the Gazetter conference, December 2006, by W. Tobler

Different ways of refereeing to geographic location can be considered aliases of each other. They differ in many attributes, such as accuracy and resolution. The main types refer either to point locations, bounded areas or routes, or can be identified by attributes.
Inter-translation between these aliases raises several problems. Several of these are discussed in my power point presentation at
http://geog.ucsb.edu/~tobler/presentations/ "Geographic location and map projections"
also see
http://www.csiss.org/classics/content/86 "Geogcoding experiment by Gould and Tobler.

As one example the CIESIN global-population-of-the-world project
http://sedac.ciesin.columbia.edu/gpw
reports on 376,499 populated areas with a global average of 46 km resolution (resolution variance not stated), containing an average of 144,000 people. Here the average resolution is defined, in km, by the square root of the country area divided by the number of units. For each of these they give a centroid point but also the polygon description, both of these in the form of latitude and longitude coordinates. They have comparable information for 24,135 urban extents and 70,558 settlement points at a resolution of 30 arc seconds. The documents available from CIESIN describe this information in more detail and discuss some of the problems associated with this type of information.

Candidate Participant Interests in Digital Gazetteer Research and Practice Workshop
December 7 – 9, 2006, Santa Barbara
Paul Veisze, GIS Manager, California Governor's Office of Emergency Services (OES)

General

- Digital gazetteers in georeferencing applications

Effective management of California emergencies depends on accurate, accessible location information, delivered in forms applicable to local needs. Emergency Managers must have the means to answer "Where is it?" and to extend that answer to the communities they serve in terms understandable by all.

- Collaboration and advancement of a research and practice agenda

OES is working with the University of California Office of the President in support of the California Hazards Institute (Rundle and others, 2006), a multi-campus initiative to leverage University resources for statewide emergency management. Placenames are key.

Core elements of gazetteers

- Placenames

I have a professional passion for the "cross-disciplinary data compression" that placenames offer: language, history, geography, sociology, technology, policy…

- Place categories

California's administrative complexity requires parity in the complexity of reference systems, particularly in the hyper-sensitive arena of public safety. I would like to launch a census of administrative names (Ranger Districts, Water Districts, and the like) that would lead to their comprehensive encoding in the USGS Geographic Names Information System (geonames.usgs.gov).

- Geospatial locations

This workshop is fertile ground for engaging debate on the merits of the National Grid (aka Military Grid Reference System) for emergency management applications.

Support Missions

- Enterprise georeferencing systems

California law mandates a Standardized Emergency Management System (SEMS). The SEMS is the model for the National Incident Management System (NIMS). I view placenames as a core element of these standardization processes.

- Geoparsing of text to derive spatial locations

Case in point: OES has just taken delivery of over 80 plans for Continuity of Operations and Continuity of Government, submitted by State Agencies. The explicit identification of alternate facility locations is central to the acceptability of the plans. Participation in this workshop would broaden my access to research and development on tools for automation of reviews and evaluation of these and related administrative documents.

- Navigation services

I offer my experience in aviation and earth imaging to the workshop community.

- Geographic information retrieval (GIR)

Emergency management (EM) represents a demanding court of engagement for GIR: systems must enable managers to ascertain what is where (in terms of vulnerable populations, evacuation routes, care and shelter resources, etc) in a hot hurry.

Selected Session Issues

- Appropriate generalization of the geospatial location

I am interested in the diversity of perspective on this…appropriate for whom, for what…

- Creation and sharing of category schemes for gazetteers

Referring to above notions of administrative naming: SEMS/NIMS need their own gazetteers: e.g. names of fire stations, police stations and their associated districts

- Integration of gazetteer data from multiple sources (crosswalks, etc)

This is among the central challenges to GIS for emergency management: OES must be able to "navigate" local, county, and regional location reference systems in order to deliver information to a wide spectrum of clients, including the Governor, federal officials, and the public.

- Interoperable gazetteer services

As above.

- Gazetteers of official toponymic authorities

As above; seeking support for processing of administrative names within the GNIS

- Place identifier tables accompanying GIS datasets

Would find applicability in the everyday work of the OES GIS Unit (7 staff, 3 cities)

Digital Gazetteers for Reference Mapping Applications

Position Statement, Digital Gazetteer Research & Practice Workshop
Santa Barbara, CA December, 2006

Howard Veregin, Ph.D.
Director, GIS Operations
Rand McNally & Company

In many ways a digital gazetteer is no different than a GIS feature class. The basic components of the gazetteer – name, type, and location – are sufficient to enable simple reference maps to be made with labeled features and symbology differentiated by feature type. Additional attribution – population, for example – allows for more complex map representations. Adding a capability to handle multiple names and multiple locations allows the map representation to be tailored to particular map specifications or product editorial guidelines. By accommodating parent-child information linking features at different levels of the geographic hierarchy, features can be selected for different types of cartographic treatment or for different naming policies.

Viewing a gazetteer as a support tool for reference map applications exposes some content and design issues that might otherwise be overlooked. This position paper discusses some of these issues with specific reference to mapping requirements for US populated places.

Is This Place Real?

Grovers Mill, NJ, is the fictional landing site of the Martian invasion in Orson Wells' *War of the Worlds* radio broadcast of 1938. It is also a small unincorporated place of about 120 people close to Princeton, with a history that goes back to the 1700s. Like Grovers Mill, most places in this country have an existence that is part real and part imaginary. Sometimes this is due to change over time – places are abandoned, or annexed by other places, or end up at the bottom of a new reservoir. Many gazetteers are littered with such "dead" entities.

Legal status also plays a role in real versus imagined existence. Incorporated places exist unambiguously because they have a legal boundary and governmental functions. Unincorporated places have neither a boundary nor a government, and some are more "real" than others. The US Census Bureau delineates CDPs (Census Designated Places) for significant population clusters that have no legal status. Some are decidedly real, such as Arlington, VA, which has a population of almost 200,000 and a name that is widely-recognized. Other CDPs, however, are for statistical purposes only. They are population clusters but have little or no local meaning or identity. For example it is doubtful that the inhabitants of the CDP called Hickam Housing, HI, have ever heard of the place.

In map production, place validity is closely tied to product specifications and editorial guidelines. For example, on a road map, all places should have some discernible physical presence on the ground that can be used for navigation. A sizable population might not qualify if it is scattered over a large area. A small population cluster might suffice if there is a cluster of houses, a fire station, or a water tower with "Welcome to Waldo, Wisconsin!" painted on its side. Attributes reflecting size or importance, such as population, are needed to identify which places take priority on a small-scale map where space is an issue, and also to tell the map user where she might find a gas station or convenience store. Historic places might be shown as a POI but not as a populated place, and statistical areas like Hickam Housing would not be shown at all.

Places do not just appear and disappear over time, but also undergo transformations. Until its incorporation in 2001, Goleta, CA, was described by the Census Bureau as a CDP with a population of over 55,000. Post-incorporation, its population dropped to around 30,000 because the footprint of the incorporated place is smaller than that of the CDP. The implications of Goleta's incorporation include possible changes to its map symbols (town dot and text size) and index entry (if population is shown). In extreme cases feature selection itself may be affected.

As these examples show, reference map applications require a rich set of feature attribution related to place type, incorporation status, historical status, local recognition, population, and so on. Since this information is expensive and difficult to collect and maintain, many existing digital gazetteers do not provide it. For example, in GNIS the basic place-type categorization includes only populated places and locales, with an optional historic designation. As such much additional research and development must occur before GNIS data can be used to support reference mapping activities.

Duplication and Apparent Duplication

In this country, place name duplication is the norm, not the exception. One reason is the relationship between intersecting administrative units, the most obvious example being municipalities (cities, towns, villages) and MCDs ("towns" and townships). Coupled with this is the existence of various official agencies, like the Census Bureau, which create areas for statistical reporting at various levels of the geographic hierarchy.

As a result, Groton, CT – an incorporated place with a population of about 10,000 – co-exists alongside Groton, CT – an MCD ("town") with a population of about 40,000. The MCD has a larger population because it covers a larger area. While these two entities have the same name (ignoring for now the issue of how the generic term "Town" is handled) they have different legal status, which is important in reference mapping due to the need to ensure that product specifications are being adhered to. If nothing else, given the different populations of these two entities, choosing one over the other would affect the dot and text size on the map.

A similar type of confusion occurs between CDPs and other delineations of unincorporated places. While CDPs are constrained along the boundaries of census enumeration units, there is no reason why this must be so. Landscan and similar population grids allow us to define population clusters more arbitrarily, which can give rise to multiple views of the same population cluster, all potentially valid relative to definitions and criteria.

A more difficult problem is the inverse – places that are truly duplicated in the database, but have not been identified as duplicates because their names are different. It is relatively easy to de-dup Goldengate, IL and Golden Gate, IL. More difficult is Branson West, MO, which is also known as Lakeview. Most places in the US have multiple place names they inherit from common usage or official status, or that have been designated by the US Postal Service or Census Bureau.

The goal of having one entry for each named place oversimplifies reality – or at least the reality of the US populated place landscape. Multiple records for named places is a common occurrence. If de-duping of records is to occur, it needs to be based on more than just name and location, as this will undoubtedly degrade the richness of the data to some degree. For reference mapping purposes it is more desirable to retain multiple records if there is some logical reason for them, such as different legal status, as long as the gazetteer provides the necessary attribution to differentiate these records, and as long as the alternate names problem is properly handled.

Geographic Hierarchies

Explicit information about hierarchical relationships between geographic features is a necessity for map production tasks. If these relationships were not available cartographers would not be able to perform even simple functions like sorting places by state to create a state-by-state index, or performing feature selection to customize symbology or naming conventions in different regions.

In the US, hierarchical relationships can be complicated, and simple spatial point-in-polygon rules often fail to provide expected results. Places can have multiple affiliations, such as Buffalo Grove, IL, which crosses the Cook/Lake county line. The same is true at the MCD level. To correctly associate places with their MCD and county "parents" it is necessary to compare polygonal representations, which can be tricky given the slivers and gaps that result from different spatial representations and levels of generalization.

Even with polygonal representations, errors can occur. In about half of the "township states" (itself a fuzzy set) there is no overlap between municipal and MCD governments. An example is Wisconsin, where municipalities and MCDs are mutually exclusive from a legal point of view. In the other half of the township states, some (but not all) of the municipalities within the state operate within territory that is also served by an MCD government. Some governmental functions are the responsibility of the municipality and some are the responsibility of the MCD. In Illinois, for example, all municipalities are within an MCD except Chicago, Cicero, and those municipalities in counties that have no MCDs. If product specs call for differentiation by governmental authority, it may be necessary to rely on other sources (e.g., the Census of Local Governments) if accurate information is not available in gazetteer format.

Or consider Hawaii, which has no legally-defined cities at all, except Honolulu, which has a single municipal government exercising control over the entire island of Oahu. The Census Bureau recognizes a CDP called Honolulu that is much smaller in size that the city/county of Honolulu. One issue is whether smaller unincorporated places on

the island of Oahu are "part of" Honolulu or not. This would affect how they would be symbolized in a product, or if they were selected for display at all.

Existing digital gazetteers take different approaches to displaying hierarchical information, from text-based explanations (GNIS) to highly structured hierarchies (Getty Thesaurus). Few gazetteers provide hierarchical information past the county level, and information is sometimes obsolete, a reflection of the expense and difficulty of maintaining data to this level of detail.

Conclusion

Reference map production is obviously just one application of gazetteer data, but it does offer some unique perspectives. A fuller understanding of the issues requires an analysis that goes beyond the US and beyond populated places alone, to encompass a broader geographical context. Still there is no question that to provide enhanced utility for detailed reference mapping purposes, digital gazetteers require flexibility, accommodation of local geographic variables, and attribute richness to try to capture the complexity of the real world.

Dr. Xiaobai Yao
Assistant Professor
Department of Geography
University of Georgia

Dear Professor Goodchild:

I am writing to apply for participation in the Digital Gazetteer Research and Practice Workshop which will be held from December 7 though December 9, 2006. With this letter, I am also applying for financial support to cover the travel and accommodation costs, in terms as described in the open call for participation that was sent to me via the UCGIS mailing list. I am an assistant professor in geography department at the University of Georgia. My primary research in the past has focused on representing and analyzing qualitative information in GIS, which fits very well with the focuses areas of the workshop. I am really excited to see the organization of this event. It is my sincere desire to be able to contribute to and benefit from the workshop.

In the past few years, I have been envisioning the "next-generation" GISs that can georeference, integrate, and analyze qualitative geographical information. I am particularly looking at qualitative information such as places names, qualitative spatial relations (near, north, in, on etc…), and qualitative modifiers (very, a little, etc.). As human beings often have incomplete and/or inexact knowledge of the environment, it is crucial for GIS services to be receptive to qualitative inputs. Qualitative descriptions such as place names are very often used in people's daily life, and they also exist in many text-based databases. Contemporary GISs do not make much use of them, neither can they represent or analyze them. With this envision of a new generation GIS, I started my research with the definition of a new concept, qualitative location, as "the reference of locations using their qualitative descriptions and/or qualitative spatial relations with other features" (Yao and Jiang 2005; Yao and Thill 2006). I first explored strategies to query and visualize qualitative locations in GIS. I then focused on the proximity spatial relations (such as near and far). A close examination of the qualitative spatial relations revealed the research challenges brought by two innate characteristics of the qualitative spatial relations: context-contingency and vagueness. These two characteristics are present not only in the proximity spatial relations, but also generally in many other qualitative descriptions and qualitative spatial relations. These characteristics bring about some very interesting and often unavoidable research issues when we try to interpret qualitative geographic descriptions. I have proposed two approaches, a neuro-fuzzy inference approach and a statistical approach (Yao and Thill 2005; 2007), to account for context factors in the translation/interpretation process. Particularly, the neuro-fuzzy inference approach deals with both context contingency and fuzziness.

Based on my prior studies, I am developing two research projects/ideas concerning the referencing and analysis of qualitative location information with the aid of GIS and other state-of-the-art technologies. The projects/ideas are in line with the first two focus areas as identified in the workshop call, and are marginally related to the third focus area. First,

I am working on theories and formal model(s) to define and support a so-called "qualitative georeferencing" (tentative) mechanism using place names and qualitative spatial relations. The mechanism is aimed to enable another level of georeferencing capability that will be based on the existing metric and discrete georeferencing mechanisms. This ongoing study is facing research issues from a wide array of perspectives such as ontology, semantics, context-contingency, and imprecision and fuzziness. While the challenges add more excitement to the research, I would definitely like to discuss and/or collaborate with other scholars who share similar interests. The workshop provides a valuable opportunity for that. Another related research idea in development is spatio-temporal data mining of qualitative descriptions. There are many text-based descriptions in large volume of databases and documents, data mining efforts could be very useful to identify spatio-temporal linkages, patterns, and relations among the qualitative geographic descriptions. A possible outcome of such an effort is the construction a spatio-temporal gazetteer with qualitative spatial-temporal linkages and relations established and embedded.

My passion in research on handling qualitative information in GIS and my desire to interact with other scholars stimulate my interests in the workshop. I believe that I can contribute to the discussions at the workshop. Furthermore, I am aware of the pioneer research on digital gazetteer at UCSB and the fact that this workshop is organized by the most extolled researchers in the field. It is clear to me that this workshop has a lot to offer and my research will definitely benefit tremendously from it. I anticipate that the event will provide great platform for individuals to shape/exchange ideas and for the community to advance the research along this line. I will be honored if I were accepted to participate in the workshop.

Yours Sincerely

Xiaobai Yao

P.S.

References

Yao, X. and J.C. Thill. 2007. Neuro-Fuzzy Modeling of Context-Contingent Proximity Relations. *Geographical Analysis: An International Journal of Theoretical Geography.* (In Press)
Yao, X. and Thill, J.C. 2006. Spatial Queries With Qualitative Locations In Spatial Information Systems. *Computers, Environment and Urban Systems.* 30(4):485-502.
Yao, X. and Thill, J.C. 2005. How Far Is Too Far - A Statistical Approach To Proximity Modeling. *Transactions in GIS*. Vol. 9(2): 157-178.
Yao, X. and Jiang, B. 2005. Visualization of Qualitative Locations. *Cartography and Geographic Information Science.* 32(4):219-229.

Open Source and Open Environment to Enrich Digital Gazetteers and Facilitating Georeferencing Processes

A position Paper to the Workshop on Digital Gazetteer Research and Practice by May Yuan, University of Oklahoma

Place names (or toponyms) are highly variant and dynamic. Multiple places may have the same name across different administrative areas at a higher order. For example, Miami in Florida is a metropolitan city, but Miami in Oklahoma is a rural town. While both places share the same spelling, their pronunciation is quite distinct: m-ai-a-mi vs. m-i-a-m-ai. Furthermore, a place may have multiple names and local variants, especially for place names that are translated from one language to another. The English spelling for China's capital city can be Peking (an earlier version) or Beijing (the current official spelling). New communities and streets are developed with new names. Existing communities and streets may experience name changes over time.

In addition to city names, names for geographic features (such as mountains and rivers) can change, and as geographic features evolve, locations and geometries associated with these names will change accordingly. Volcanic eruptions or landslides can quickly alter the correspondent morphological and geometrical (shape and spatial extent) associated with these geographic features. New toponyms are given to new geographic features developed by natural or man-made processes (e.g. lagoons, retention ponds). Besides names, relationships among places and/or geographic features can change as well. These relationships may be containmentship, intersection, distance, and other non-spatial or spatial cases. The highly variant and dynamic nature of place names makes it challenging to build a comprehensive digital gazetteer of the world by any one or group of organizations. Hence, the position paper promotes the use of open source information in an open environment to enrich digital gazetteers that take the advantages of rich information from different places, and broad-based local knowledge on the World Wide Web.

Open source information can be on-line or off-line. Off-line open sources are documents or records open for inquires and browsing, such as unclassified government documents, newspaper, books, and other academic literature, etc. While there may be charges to access these off-line open source materials, the information is in general available for the public. The growth of cyberinfrastructure democratizes further publications and dissemination of facts, information, and knowledge. Quality of internet posting, however, varies, but a range of mechanisms has been used to build credibility and reliability. E-commerce and wikipedia are two examples of great success. E-commerce takes user feedbacks to build a reputation for sellers, buyers, or products. Wikipedia, on the other, provides a collaboratory open environment in the cyberspace to build the most comprehensive encyclopedia with the broadest and most diverse author communities. Both e-commerce (reputation established by peer feedbacks) and wikipedia (broad-based authorship) models offer new thoughts to the use of open source and open environment approaches to enriching digital gazetteers. A board-based authorship allows extensive and intensive incorporation of local knowledge that is critical to address the variant and dynamic nature of place names, while peer feedback mechanisms provide a measure of credibility to the authorship and local knowledge. Moreover, a board-based authorship promotes the opportunity to supply historical and geographic contexts to individual place names, most of which have historical, cultural, geographical, or social traces.

Attempts to a broad-based authorship and peer-feedback mechanisms challenge modeling of gazetteer data and meta-data. The traditional alphabet ordering approach cannot effectively handle frequent updates and added information, such as authorship, credibility, spatial footprints (including vector, raster, and imagery data) and versioning, place name time (the time when the place is in use), transaction time (the time when the place name is entered to the gazetteer), and the temporal lineage of place names. Design of such a gazetteer information system shall consider a database with three domains of semantics (e.g. place names, authors, evidence, and contexts), time (e.g. valid time of place names, transaction time of data entries), and space (e.g. spatial footprints, spatial relationships). A semantic object of a place name may be linked to multiple temporal objects of valid time or transaction time, and then linked to multiple spatial objects of footprints and relationships. If a place name corresponds to more than one location, then the place name will be linked to multiple spatial footprints. Similarly, when a place expands (through urban sprawl, for example), the semantic object of a place name will be linked to multiple temporal objects of valid time and then to multiple spatial objects to represent transitions in spatial extents over time. On the other hand, a spatial object may be linked to multiple temporal objects and then to multiple semantic objects to represent a location may have multiple place names over time.

When implemented, advanced search engines are possible to extend place name queries from "where is place name X" and "what is at the place Y" to "How many places have the place name X" and "How has the place changed its name over the years and what is the context for the name changes?" Searches initiated from the spatial domain seek all place names that have been used for a location at the best knowledge of the system. Searches initiated from the semantic domain inquire all locations where a place name has been used to reference these locations. When temporal objects are referenced, searches can be extended to transitions of place names at a location, and an increasing or a decreasing use of a place name over space and time. Subsequent analysis can be done to examine historical and geographic implications for the stability of a place name. Are there regions experiencing common place name changes? Did certain place names become popular in temporal periods or geographic regions? Are certain place names particularly transitory? Are there certain place names commonly embedded special religious, ethnological, social, or functional meanings?

Furthermore, the enriched gazetteer will enhance georeferencing. The gazetteer-based geoferecing process is no longer merely a match between place names and geospatial footprints. Temporal references will indicate how place names evolve at a location or for a geographic feature as well as added historical and geographic contexts to facilitate a better understanding of a place than just its name and geospatial footprint. In addition, the added temporal lineages and geographic context can be used to improve the accuracy of georeferencing, especially when a place name is used by multiple locations or a location has more than one place name. Temporal lineages and historical/geographic contexts help narrow in ambiguity, relate the place to its former or later place names, and associate the place to other place names in its surroundings.

In sum, open source and open environment can offer great promises to expand the current approach to develop digital gazetteers. The expansion can address the challenges of variant and dynamic nature of gazetteers. Advanced place name searches and context analysis of place names can greatly improve the functionality and usefulness of digital gazetteers to facilitate our understanding of places over the world. As the term "place" emphasizes the geographic context of a location, gazetteers built with a broad authorship and local knowledge address directly the

essence of semantic, temporal, spatial components of a place and help us to understand places with historical, geographic, and social contexts.

International Workshop on

Digital Gazetteer Research & Practice

7-9 December 2006

Convened by	The workshop takes place at
National Center for Geographic Information and Analysis (NCGIA) University of California, Santa Barbara and Redlands Institute, Redlands, California Sponsored by the National Geospatial-Intelligence Agency (NGA)	Upham Hotel 1404 De La Vina St, Santa Barbara, CA 93101

AGENDA

Thursday 12/7

All day – Arrivals; poster setup

5:30 pm - Welcome Reception (hosted bar, enough eats to hold you until dinner)
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

Friday 12/8

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

10:30 - 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

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

3:30 - 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

Saturday 12/9

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

10:30 - 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

1:30 - 2:30 Breakout Session I

2:30 - 3:00 Reports from breakout groups

3:30 - 4:30 Breakout Session II

4:30 - 5:00 Reports from breakout groups

5:00 - 6:00 Concluding discussion and future directions