INTRODUCTION TO DIGITAL GAZETTEER RESEARCH

Michael F. Goodchild, National Center for Geographic Information and Analysis, and Department of Geography, University of California, Santa Barbara, CA 93106-4060, USA. Phone +1 805 893 8049, FAX +1 805 893 3146, Email good@geog.ucsb.edu

Linda L. Hill, National Center for Geographic Information and Analysis, and Department of Geography, University of California, Santa Barbara, CA 93106-4060, USA. Email l6hill@verizon.net

ABSTRACT

Digital gazetteers provide information on named features, linking the feature's name with its location and its type. They have been growing in importance recently as the basis of a range of services, including wayfinding, georeferencing, and intelligence. This introduction to the following collection of five research papers expands on contemporary applications of digital gazetteers, explores the issues associated with each of the three types of information, and defines three broad areas of research: the *components* of gazetteers; the *process* by which places are named and evolve; and the issues of *interoperability* between digital gazetteers. Each area is represented by at least one paper in the collection. Digital gazetteers increasingly form the interface between the informal discourse of humans and the formal world of geographic information science.

1. BACKGROUND

The building and application of digital gazetteers, which organize knowledge and details about named places, is a growing area of research that necessarily involves cross-domain issues from fields such as spatial cognition, geographic information science, social history, computer science, and geographic information retrieval. Gazetteer-based services are being developed and deployed in such specific fields as public health, natural history data management, cultural history, and automated georeferencing of text (geoparsing). Gazetteers started out as reference books that documented the names and accepted spellings of places (toponyms) -- for example, standardized placename lists for newspaper articles. In these gazetteers, contextual and descriptive information about the places was included as needed. Some of these developed into sources of authoritative names which were used widely, and the organizations that prepared them, mostly for governmental purposes, were and are known as toponymic authorities. Such authorities exist for many countries, as well as for constituent states, provinces, and even local areas. In the U.S., for example, the Board on Geographic Names was established in 1890 with the objective of standardizing the use of placenames. Thus the emphasis in all of these efforts was on the naming of places.

When computerized mapping software emerged, placenames were treated as a *names layer*, often structured as a table containing placenames that was linked to named features for the purposes of labeling. In this case, the emphasis is on representing the geographic location rather than on the naming. In the 1990s, research projects such as the Alexandria Digital Library (ADL) project at the University of California, Santa Barbara demonstrated the key role that gazetteers play in geographically enabled information

management and retrieval systems. It became clear that gazetteers play a vital role in information systems because they encode relationships not only between placenames and geographic locations but also between these elements of place description and the type of place (e.g., *river*). As the importance of digital gazetteers in information systems grew, it became apparent that the element of time and the relationships between named places are also and perhaps equally important for some applications.

In the past few years the naming and description of places has taken on new meaning and importance with the growth of user-generated content on the Web. Sites such as Wikipedia, Wikimapia, and Flickr allow private citizens to create and integrate descriptions of interesting places and to build links between them using standard codes known as *geo-tags*. The number of places described in Wikimapia already matches the number of authoritatively named places in gazetteers, and the number of georeferenced photographs in Flickr exceeds that number by at least an order of magnitude. This populist approach to geographic information creation has been termed *neo-geography* (Turner 2006), and it offers an interesting alternative to the more traditional system of naming authority. Proponents would argue that the elements of *collective intelligence* or *crowd-sourcing* that are present in these activities, in which contributors are able to challenge or edit the earlier contributions of others, is the modern equivalent of the process of consensus that the naming authorities have traditionally relied on and managed.

2. COMPONENTS OF DIGITAL GAZETTEERS

Digital gazetteers contain structured information about *named places*, where *place* is defined simply as a geographic location and more expansively as a geographic location that has been identified and referenced as a social construct. It is useful to distinguish how *place* differs from *feature* in usage, even though the terms are often used interchangeably. Features are distinct physical elements or objects in the landscape, such as mountains, rivers, and buildings, for which definite geographic locations and boundaries can be given. Administrative areas are often considered to be features because they also have clearly identifiable locations and boundaries that are established by fiat (Smith and Varzi 2000) distinguish between *fiat* and *bona fide* boundaries). The meaning of *place* includes features and also locations with vague positions and boundaries (Burrough and Frank 1996), for example the Rocky Mountains and southern France. A place can also be described as something like "down by the river" that has meaning related to experience -- e.g., memories of riverside picnics and lyrics of a song -- and "I'm going downtown". These are examples of references to generic places of certain types (rivers and downtowns); they are references to particular locations only when they are identified as being *the* river or downtown within a particular spatial context.

Gazetteers hold structured information about *named places* that have a particular geographic location; that is, the subset of places that have acquired *proper* or authoritative names (e.g., London) or sections of places with proper names (e.g., east London). The characteristics of these names are undefined. They do not have to designate a particular place uniquely (i.e., the same name can denote several different places) nor do they have to be officially sanctioned. They may be "cute"; for example, the "Study

Hall" which is a bar in a student neighborhood. There may be several names for the same place, varying by date, language, spelling, and official versus colloquial. They may be assigned for special purposes (e.g., Deep Sea Drilling, Leg 42, Hole 378) and be understood and used only in specialized domains.

Because of developments in the past ten years, including research into georeferenced digital libraries by the Alexandria Digital Library (ADL) project at the University of California, Santa Barbara, standards development by the TC 211 committee of the International Organization for Standardization (ISO), and protocol development by the Open Geospatial Consortium, there is general agreement on the structure of digital gazetteers, the attributes of place descriptions in gazetteers, and protocols to support distributed searching of gazetteers created for different purposes. The minimum required elements of a place description are represented by the tuple *N,F,T* where *N* is at least one name, *F* is at least one representation of geographic location according to a mathematical framework (i.e., a footprint), and *T* is a least one type (category, class) drawn from a typing scheme (i.e., a system of feature classification). Since places themselves come and go with the passage of time as well as the elements of place description, gazetteers must also incorporate date ranges in multiple ways in any model of named place description.

A further gazetteer characteristic is the degree to which spatial and temporal representations can be generalized and still meet application requirements. When the purpose of the footprint is to disambiguate one river from another within a gazetteer, then

point locations in the mouths of the rivers are sufficient; this is often the case with toponymic authority files. Simple point locations also suffice for orienting map displays -- for moving the map view over to that location and zooming in. For geographic information retrieval (GIR), using minimum bounding rectangles (MBRs) to represent geographic coverage of a place is both computationally easy for overlap analysis and sufficiently effective with sophisticated matching methods to result in satisfactory performance. For places with naturally vague boundaries (e.g., the Rocky Mountains and Southern France), footprint generalization provides coverage information without forcing an arbitrary boundary line.

3. TOWARDS A RESEARCH AGENDA

Digital gazetteers are an increasingly important form of geographic information, representing in effect the interface between the informal world of human discourse (people mostly use names to refer to parts of the world) and the formal world of geospatial technology, with its scientifically rigorous systems of georeferencing. Yet the names layer has occupied a somewhat uncertain position, and does not appear for example in the seven *framework* themes of the U.S. Federal Geographic Data Committee, though certain types of features may appear in the administrative, hydrographic, and transportation themes. *Toponymy*, the study of placenames, is a recognized area of scholarship but hardly ranks among the geographic information sciences. Nevertheless there are very substantial problems in digital gazetteers that deserve scientific study.

In late 2006 we organized a three-day specialist meeting in Santa Barbara on the topic of Digital Gazetteer Research and Practice, to bring together a wide range of researchers and practitioners who were working with the definition, modeling, and application of gazetteers and gazetteer services, to present and discuss current activities and short- and long-term research and development directions. The 43 participants, from seven countries in addition to the U.S. and representing academia, industry, and government, were selected by invitation and from responses to an open call. Our final report of the meeting (Goodchild and Hill 2007) is available on-line, along with the presentations, position papers of participants, and much related material, at the meeting Web site http://ncgia.ucsb.edu/projects/nga/. At the meeting there was strong interest in a special journal issue that would showcase examples of gazetteer research; this issue is the result of a process of submission and peer review that began at that time.

4. INTRODUCTION TO THE PAPERS

We structured the specialist meeting under three headings:

1. Components of gazetteer services

The three core elements of gazetteers – placenames, place categories, and geospatial locations – support the translation between informal georeferencing using placenames ("Santa Barbara") and place categories ("city") and the formal georeferencing of mathematical schemes (e.g., longitude and latitude coordinate systems). These elements plus explicit relationships between named geographic places and the identification of time frames for places and their characteristics are the fundamental components of digital gazetteers. Within the context of gazetteer services, such as support for enterprise

georeferencing systems, geoparsing of text to derive spatial locations, navigation services, and support for geographic information retrieval (GIR), the complexities of each of these components challenge the collection and use of gazetteer data.

2. Georeferencing as a process

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

3. Interoperable gazetteer services

Gazetteer data exist in many independent sources that are often dissimilar in construction and content, including:

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

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

The five papers in this special issue straddle the three topics, and have been organized in corresponding sequence. The following paragraphs introduce the papers, and explain their relationship to the larger gazetteer research agenda.

Most gazetteer services to date have been built on existing digital files provided by naming authorities, and thus cannot recognize vernacular names that refer to features without official name recognition. Moreover gazetteers that represent features using a single point, MBR, or other highly generalized boundary are clearly less useful for certain types of information retrieval. Thus there is substantial interest in the development of methods that can readily provide detailed geographic footprints, when existing footprints are either non-existent or highly generalized. The paper by Jones, Purves, Clough, and Joho describes a novel and elegant approach to this problem that mines the use of a placename in Web pages, and returns a probabilistic estimate of its footprint.

The second paper in the gazetteer services area also adopts a probabilistic approach, developing methods to model uncertainties in vague geographic referents. The spatial interpretation of textual georeferencing involves the use of gazetteers to recognize placenames in text strings. Once a possible placename is identified, surrounding text is used to verify that the name is a place reference (e.g., Clinton, Arkansas rather than Bill or Hillary Clinton), to identify which “Springfield” is referenced, and to find phrases indicating an offset from the place (e.g., five miles south of Bakersfield). The paper by Guo, Liu, and Wieczorek presents methods developed to estimate the geospatial location

based on such references found in the descriptions of natural history museum collections. This research illustrates the great potential of, and challenges to the spatial analysis of the wealth of historical specimen data held by the world's natural history museums. The paper builds on previous work on modeling uncertainties in geographic information, applying it to the specific needs of placename references. A powerful software package is described that implements the techniques under a straightforward user interface.

The second topic, the process of georeferencing, is clearly best studied from a temporal perspective. While several participants at the specialist meeting found this topic to be of great interest, it is represented in this collection by just one paper, by Mostern and Johnson. In it, the authors reorient our thinking about gazetteers by introducing a more time-centric gazetteer model based on spatio-temporal events. They argue for an approach that recognizes that the concept of events extends to the places themselves, which, as human constructs, have starting and ending times and associated events that create and change them through time. From the perspective of cultural history, they see such an event model as mirroring the methodology of historians as they piece together the multiple sources of evidence for past events and locations.

The final two papers represent the third topic, of gazetteer interoperability. Gazetteers are collections of *gazetteer entries*, each of which contains information about a particular named place. Ideally, a particular gazetteer will contain only one entry per place and this goal raises issues of how and with what certainty it is possible to tell when two pieces of information are about the same place. The difficulties arise because the same place can

have names that differ to some degree, and can have footprints that vary by type (e.g., point, MBR, or polygon), date, accuracy, and specificity, and which probably are assigned to different categories depending on the typing practices applied. The paper by Hastings addresses this conflation issue, presenting a detailed strategy and associated evaluation metrics that can be used to achieve conflation between multiple, conflicting gazetteer entries.

The final paper by Janowicz and Kessler focuses on the type element of a gazetteer tuple, and discusses the issue of interoperability with respect to this element. The topic of semantic interoperability, or the sharing of meaning across disparate systems and data sets, is particularly challenging and has been the subject of much recent research. The paper argues for a more rigorous, theoretically grounded approach to defining type classifications that is as a result more readily compatible with the ontological systems that have been constructed in recent years to facilitate semantic interoperability. Much work remains to be done in this area, which is one of the most important challenges of 21st Century computing.

5. CLOSING COMMENTS

This issue of the *International Journal of Geographic Information Science* (IJGIS) appears as the scientific community begins to grapple with the fundamental and applied issues raised by digital gazetteers and the gazetteer models and standards developed in the 1990s, and to respond to the very rapid recent developments in user-generated Web content and gazetteer-based services. At some point in the future, it may be possible to

submit a query such as “find an orange grove five miles north of Bakersfield” to a portal and to obtain a probability density function that has been informed by all of the many sources of relevant information distributed over the Web. Before that can happen, a number of issues of semantic interoperability, uncertainty, and conflation must be resolved by research, and the results implemented. On a global level, a related query such as “what is the name of the large river that flows north out of Mongolia” will require solutions to additional problems, including interoperability between placenames in different scripts and languages. Phonic representations of placenames is another interesting area of potential research that may help in achieving interoperability at the global level.

Inclosing this introduction, we return to two earlier themes. First, despite the general lack of scientific interest in placenames, it is evident to us that the names layer is of vital and increasing significance in GIScience, because of its role in supporting the general user’s interaction with geographic information technologies. Second, and related to the first, is the essential role that gazetteers play in facilitating communication between the informal, idiosyncratic, and often vernacular world of human discourse and the formal, scientific, and hopefully interoperable world of GIS. In that sense, gazetteers offer one of the keys to a more human-oriented paradigm of user interaction. We hope that the papers of this special issue, and the results of the specialist meeting, will stimulate further research and development in this increasingly important area.

ACKNOWLEDGMENTS

We thank the 43 participants at the specialist meeting for stimulating discussion, doctoral students Jordan Hastings and staff member Guylene Gadal for their assistance, the 37 reviewers who provided helpful comments on the papers submitted for this special issue, and the National Geospatial-Intelligence Agency for financial support.

REFERENCES

BURROUGH, P.A. and FRANK, A.U. (Eds), 1996, *Geographic Objects with Indeterminate Boundaries* (Bristol, PA: Taylor and Francis).

GOODCHILD, M.F. and HILL, L.L., 2007, *Summary Report: Digital Gazetteer Research and Practice Workshop* (Santa Barbara, CA: National Center for Geographic Information and Analysis).

SMITH, B. and VARZI, A.C. 2000, Fiat and *bona fide* boundaries. *Philosophy and Phenomenological Research*, **60**(2), pp. 401-420.

TURNER, A., 2006, *Introduction to Neogeography* (Sebastopol, CA: O'Reilly).