4 Exploring the Use of Gazetteers and Geocoders for the Analysis and Interpretation of a Dynamically Changing World

Daniel W. Goldberg, John P. Wilson, and Craig A. Knoblock

CONTENTS

INTRODUCTION

The observation that the world consists of a series of interesting and dynamic places is easily demonstrated by the following examples. If the region of interest were South Asia in the middle of the 19th century, an intelligence analyst focused on present-day Bangladesh might be interested in the partition of India and subsequent transformation of East Pakistan into Bangladesh, capturing what countries were involved (Burma, India, West Pakistan?) and how they have evolved over time. Hornsby (2001) used this example to (1) illustrate how the views of an object's evolution can be refined or coarsened and (2) describe a new set of temporal zoom operators to model these shifts in granularity. At the finest granularity, a data repository would need to capture all known past space-time events concerned with an entity, including relations between different entities or cross-links, such as the link between the dissolving of British India and the creation of independent India, Bangladesh, and Pakistan (Hornsby 2001). Clearly, the changes in objects and space-time events that spawn them are ongoing (witness the recent transformation of Bombay into Mumbai superimposed on a rapidly expanding metropolitan region — the latter implies that the spatial footprint requires continuous updating), and one can immediately see the great challenges incurred in the design and population of a gazetteer database that incorporates these changes. Some of these details cannot be known with certainty, indicating the need for support of uncertain queries as well.

Plewe (2002) outlined the beginnings of a theory of conceptual database design for documenting the nature of uncertainty in historical geographic information (herein termed geo-historical information). He relied on Frank's (2001, p. 670) assertion that we seem to structure our world around objects or entities, and described how these entities occur at particular places (locations), at particular times (life spans), and with particular attribute values (descriptions). These distinctions are more often fuzzy (i.e., intertwined) rather than crisp, given that multiple aspects can and are likely to interact in each manifestation. Hence, the location and description can each vary over time (within the overall time span) so that different points in space and attribute values are valid parts of the extent at various times. Plewe (2002) proposed the Uncertain Temporal Entity Model to describe the variety of causes, types, and forms of uncertainty present in geo-historical information and documented how it could be used to model uncertainty in digital stores such as the gazetteer databases discussed herein.

This discussion illustrates why the intelligence analyst may need to reconcile many different representations of the world drawn from many different sources in order to understand and anticipate a series of evolving or unfolding events and movements. The remainder of this chapter explores the latest developments of two sets of interrelated tools that help facilitate the analysis and interpretation of our dynamically changing world and the similarly dynamic data that describes it: the gazetteer and the geocoder.

MODELING THE CHANGING WORLD — THE GAZETTEER

The gazetteer, long known for its ability to translate a named geographic place into a geospatial footprint, has become the focus of a great deal of scientific research in

recent times (e.g., Wang and
underlying data sources, da
characteristics for geographi
and nonspatial (e.g., its type
teers. The gazetteers of old,
and/or accuracy, and incom
highly accurate, and comple
2007, GeoNames.org 2007).
ment of models capable of c
four subsections.

Background

The three gazetteers most c
(Alexandria Digital Library
and GNIS (U.S. National C
milestones in the developme
often leave those attempting
aged by the lack of detail. Ur
examples, first by the U.S. g
the University of California a
framework consisting of the
this work, the ADL team sho
issues encountered integratin
questions about ontology ali
the data structure of the gaze
Zheng 1999; Hill 2000, 200
focused on the integration an
and Hill's (2000) satisficing c
power, and time were all lim
included as gazetteer entries.

If we view the present st
we can see that gazetteers a
tial workflows. Perhaps most
are based have been concret
consist, at a minimum, of a t
graphic footprint (Hill 2000).
atomic data structure, the mar
Used as a translator from te
geographic-classification syst
model (Agouris et. al 2000), it
tions and geospatial functions
(Alexandria Digital Library 2

The aforementioned defi
ADL Gazetteer Content Stand
tation for the gazetteer featur

recent times (e.g., Wang and Ge 2006, Goodchild and Hill 2007). Advances in the underlying data sources, data gathering tools, and operations necessary to derive characteristics for geographic features from the data, both spatial (e.g., its footprint) and nonspatial (e.g., its type and name) are changing the contents and uses of gazetteers. The gazetteers of old, characterized by their rigid construction, low resolution and/or accuracy, and incompleteness, have begun to give way to high-resolution, highly accurate, and complete user-specific/centric gazetteers (e.g., WikiMapia.org 2007, GeoNames.org 2007). These changes and their repercussions for the development of models capable of describing the changing world are explored in the next four subsections.

Background

The three gazetteers most commonly used in research and practice — the ADL (Alexandria Digital Library 2007a), GNS (U.S. Board on Geographic Names 2006), and GNIS (U.S. National Geospatial-Intelligence Agency 2006) represent major milestones in the development of the gazetteer as a useful spatial-data resource but often leave those attempting to use them for high-resolution applications discouraged by the lack of detail. Undoubtedly, great effort was expended to generate these examples, first by the U.S. government to create the GNS and GNIS, and later by the University of California at Santa Barbara ADL team to create a unified gazetteer framework consisting of the GNS and GNIS as well as other data sources. Through this work, the ADL team showed that it is indeed possible to overcome many of the issues encountered integrating heterogeneous gazetteers — their publications tackle questions about ontology alignment as well as those about storage, retrieval, and the data structure of the gazetteer (Frew et al. 1998, 2000; Hill et al. 1999; Hill and Zheng 1999; Hill 2000, 2006). Understandably, the development of the ADL was focused on the integration and modeling of gazetteer data for information retrieval, and Hill's (2000) satisficing condition proved useful at the time, as resources, manpower, and time were all limited, and choices had to be made as to what should be included as gazetteer entries.

If we view the present state-of-the-art gazetteers in the broader context of GIS, we can see that gazetteers are now considered to be valuable components of spatial workflows. Perhaps most important, the three axes on which gazetteer features are based have been concretely defined; that is, that a valid gazetteer entry must consist, at a minimum, of a toponym (name), a geographic feature type, and a geographic footprint (Hill 2000). By defining the required attributes of this most basic atomic data structure, the many roles of the gazetteer could then begin to take shape. Used as a translator from textual terms to spatial footprints (Goodchild 1999), a geographic-classification system (Wang and Ge 2006), and a stand-alone spatial data model (Agouris et. al 2000), it now forms the basis for numerous geographic applications and geospatial functions spanning research communities from digital libraries (Alexandria Digital Library 2007a) to health (Dugandzic et al. 2006).

The aforementioned definition of a gazetteer continues to evolve. The current ADL Gazetteer Content Standard (GCS), for example, allows for temporal representation for the gazetteer feature itself, as well as the components of each of the axes

that describe it (Alexandria Digital Library 2007b) based in part on work undertaken by the Electronic Cultural Atlas Initiative (ECAI; Buckland and Lancaster 2004) that demonstrated shortcomings in the original definition. The result of this effort is that the changing nature of geographic places can be represented across time periods in terms of changes to their names, footprints, types, and relationships in both the spatial and typographic hierarchies in which they are contained. This fundamental enhancement is responsible for enabling the gazetteer data structure to truly capture and represent the dynamism of geographic features.

Many fundamental challenges remain to be solved, however. First and foremost are the notions of completeness and accuracy (Van Rijsbergen 1979), both at the atomic (individual entry) and holistic (entire gazetteer) levels (Smith and Mann 2003, Doerr and Papagelis 2006). Atomically, completeness refers to the amount of knowledge about each axis of an entry (e.g., Are all possible names for an entry present?), while accuracy refers to the correctness of knowledge about each axis (e.g., Are all the types associated with a feature correct?). The status of these measures at the atomic level is important because they will directly impact applications that use individual features in analysis. Similarly at the holistic level, completeness (How much of the real world is represented by entries in the gazetteer?) and accuracy (How well does the descriptive data in the gazetteer represent the real world?) measures are required for evaluating individual gazetteers and for applications working with aggregate or large data sets. The challenges here are twofold: first, developing methods to assess the completeness and accuracy; and second, developing methods to actually increase both the completeness and accuracy of individual entries and the gazetteer as a whole.

A second issue is the integration of heterogeneous gazetteers. The ADL experiments proved that it is indeed possible to integrate gazetteers, and that the results are extremely valuable. They also showed, however, that integration presents an arduous, difficult, and time-consuming task, with the approach taken entirely unable to handle tens to hundreds of gazetteers. Automated means to achieve feature ontology alignment are being developed (e.g., Doerr 2001), and more will be required in order to scale to a situation where all existing field-specific gazetteers are combined into a single coherent framework. Aside from simply mapping between feature typing schemes, integration requires reasoning to determine and resolve conflicts between entries from different gazetteers. Should this be done at the integration level while the gazetteers are being merged, when integrating results sets from distributed gazetteers, or by the end user? This and other related questions will need to be addressed before very large-scale integration attempts are made, possibly using prior research from related fields (e.g., Naumann and Haeussler 2002).

A third issue relates to the handling of temporal information for geographic features. At present, it is unclear if the representation of time for geographic features is best separated from the three original axes of the gazetteer into its own, fourth axis. The current ADL GCS maintains temporal information with each axis, rightly illustrating that data about the geographic feature have a temporal extent. But we must remember that one of the main contributions of the original ADL work was the separation of the FTT structure from the gazetteer structure, explicitly indicating that while related, the two are unique and serve separate functions. The same can be said

of the gazetteer and time pe
a common temporal ontolog
temporal data for all attribut
these temporal thesaurii bec

In addition to general q
cific issues arise when co
Commercial buildings chan
ments are constructed seen
regions of all resolutions ch
status attributes (i.e., curre
ADL CGS (Alexandria Dig
detailed representation or f
indeed be used, but are in p
specific historically orientec
questions arise, such as: (1
a single instant in time or t
extents are/can be stated pr
how to handle overlaps in te
temporal footprints.

A final, and perhaps ec
geography a gazetteer shou
their representations (Lam a
etteers can best be describe
geographic areas such as pc
most common occurrence,
tion gazetteers containing a
every inch of geography in
(Axelrod 2003).

The availability of new
exploit it to its fullest poten
gazetteers. The next three s
used to overcome portions o

Changing Methods for (

Like never before, geograph
available online. Whether it
containing geographic featu
online directories and phon
robust and mature enough to
sources (e.g., Web pages) ir
extraction technology has b
2002 for a review of existin
geographic domain (Himme
entists and researchers can
about a geographic area of i

of the gazetteer and time periods (Buckland and Lancaster 2004). Due to the lack of a common temporal ontology or thesaurus, however, the easiest method is to include temporal data for all attributes as does the ADL CGS, albeit with a stub in place should these temporal thesaurii become available (Alexandria Digital Library 2007b).

In addition to general questions about the structure of time in a gazetteer, specific issues arise when considering the dynamic nature of geographic locations. Commercial buildings change tenants often (and therefore uses/types), new developments are constructed seemingly overnight, and geographic and/or administrative regions of all resolutions change names and boundaries sporadically. The temporal status attributes (i.e., current, former, and proposed) used most commonly in the ADL CGS (Alexandria Digital Library 2007b) clearly indicate the need for a more detailed representation or format. Time periods and other textual descriptions can indeed be used, but are in practice, infrequently encountered, with the exception of specific historically oriented gazetteers (e.g., Buckland and Lancaster 2004). Further questions arise, such as: (1) the circumstances when the appropriateness of either a single instant in time or time spans should be maintained; (2) whether temporal extents are/can be stated precisely or a fuzzy attribute scheme should be used; (3) how to handle overlaps in temporal extents; and (4) how to deal with uncertainty in temporal footprints.

A final, and perhaps equally contentious issue is what scale (or resolution) of geography a gazetteer should characterize, both in terms of types of features and their representations (Lam and Quattrochi 1992, Agarwal 2004). Most existing gazetteers can best be described as low resolution, with entries that encompass large geographic areas such as populated places, mountains, and so on, being by far the most common occurrence, represented as single geographic points. Higher resolution gazetteers containing a name, type, time span, and encompassing a footprint for every inch of geography in a region may be needed for some applications, however (Axelrod 2003).

The availability of new types of data and the creation of appropriate tools to exploit it to its fullest potential are changing both the contents and applications of gazetteers. The next three subsections show how emerging technologies are being used to overcome portions of the first, third, and fourth issues discussed above.

Changing Methods for Gazetteer Creation

Like never before, geographic information in many forms and formats is becoming available online. Whether it is traditional GIS data sets such as raster or vector files containing geographic features, or newly emerging nontraditional data types such as online directories and phone books, information extraction tools are fast becoming robust and mature enough to reliably and accurately turn semistructured information sources (e.g., Web pages) into structured databaselike resources. While this Web-extraction technology has been around and in use for some time (see Laender et al. 2002 for a review of existing tools), only recently has it begun to be applied to the geographic domain (Himmelstein 2005, Laender et al. 2005). With these tools, scientists and researchers can now automatically harvest vast amounts of information about a geographic area of interest simply by creating "Web agents" or "wrappers"

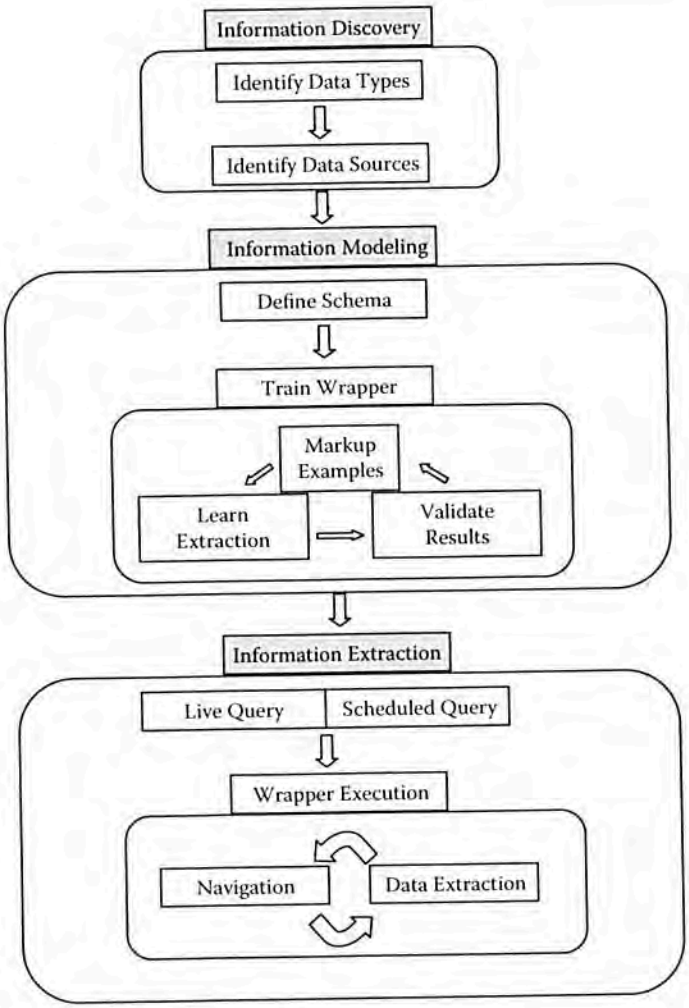


FIGURE 4.1 Typical wrapper creation and execution workflow.

around data sources that previously had no automated query ability. A typical, simplistic wrapper creation and execution workflow, including the three common stages necessary for information extraction (discovery, modeling, and extraction), are displayed in Figure 4.1 along with their component parts.

The information-discovery process begins by identifying the data types to be extracted and sources that can be used to produce them. In the modeling process, one first defines the schema for the information one wishes to extract (including the between Web page navigation structure if necessary). Through the iterative process of marking up examples of data to be extracted (training data), learning the extraction rules, and validating the results (on separate test data), the agent develops the most general and correct syntactic rules that can be applied to the pages to extract the desired information. It should be noted that validation (test) examples that fail to

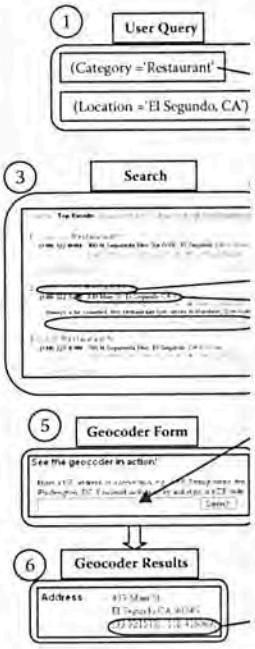


FIGURE 4.2 Example workf

extract properly given a cur
into the training set before
ing learned a sufficient set
tion engine (e.g., Barish and
live queries) or offline (i.e.,
the application and user ne
(navigation and/or extractio
by synchronously or asynch
performing database writes
just provided is but a rudir
ing and execution. Each co
an expanding literature, in
(simplifying the markup p
and their descriptions of re
consult Laender et al. (200

In Figure 4.2, we displ
in Table 4.1 to quickly prod
varying subtypes of "Rest
in turn a subclass of "Build
agents, one for each of the

The first agent will wra
and take as parameters the
(denoted by the number 1

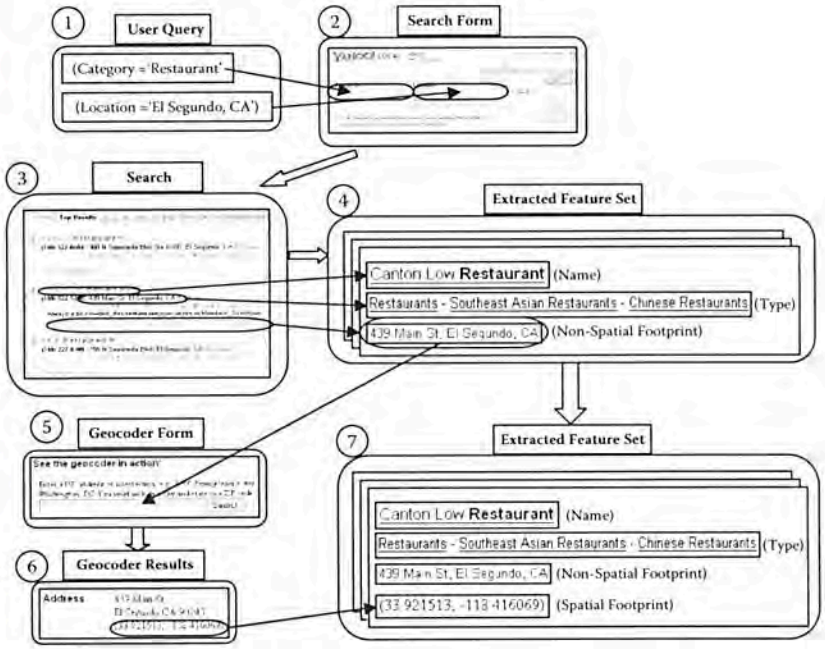


FIGURE 4.2 Example workflow for building a simple restaurant gazetteer.

extract properly given a current set of rules are typically the best choice to integrate into the training set before learning is reattempted (Knoblock et al. 2001). Having learned a sufficient set of extraction rules, an agent can be run using an execution engine (e.g., Barish and Knoblock 2005) in either an online (i.e., responding to live queries) or offline (i.e., being invoked through a scheduler) mode depending on the application and user needs. This execution engine runs each part of the agent (navigation and/or extraction) and delivers the output in a user-defined manner (e.g., by synchronously or asynchronously streaming nonpersistent XML or persistently performing database writes). It should further be noted that the description we have just provided is but a rudimentary introduction to the general idea of agent building and execution. Each component of the process is an active area of research with an expanding literature, including such topics as automatic schema identification (simplifying the markup process) and wrapper re-induction (for when the sources and their descriptions of real world features change). The interested reader should consult Laender et al. (2002) and Turmo et al. (2006).

In Figure 4.2, we display how these tools can be applied to the Web sites listed in Table 4.1 to quickly produce a simple gazetteer containing geographic features of varying subtypes of "Restaurant," a subclass of "Commercial Building," which is in turn a subclass of "Building." This workflow will require the construction of two agents, one for each of the websites in Table 4.1.

The first agent will wrap the http://local.yahoo.com (Yahoo! Inc. 2006) Web site and take as parameters the category and location of interest. Stage 1 of the workflow (denoted by the number 1 in a circle), indicates the user entering a category (i.e.,

TABLE 4.1
Data Sources Used for Creating a Gazetteer of "Restaurants"

Data Source	Parameters	Feature Information Returned
http://local.yahoo.com	Category, location	Types, names, addresses
http://geocoder.us	Address	Footprints

feature type) and a location. A first wrapper will be created around the input search form (stage 2), and the parameters supplied as agent inputs will be submitted to the site for processing. The result of this form submission will be a page containing a list of restaurants (stage 3) that will also be wrapped to extract the list of restaurants with their names, subtypes, and addresses (stage 4). These attributes contain two (toponym and type) of the three axes required to be considered a true gazetteer feature (toponym, type, and footprint). The footprint will be obtained from the second site through the process of geocoding (this process is explained in some detail later in this chapter). Obviously, the exact start and end times for when the information obtained was valid are not explicitly available, but we can associate a time period to the sources from which the data are obtained. The simplest and safest method is to assume the most conservative estimate possible, that the sources used represent the "current" state of the world and information about the features contained therein. Therefore, at this point all that is known about the temporal extents of the attributes describing these axes is that they are "current," in the parlance of the ADL GCS (Alexandria Digital Library 2007b).

A second agent will be created around the http://geocode.us site, and the first wrapper will be created around the input form located at http://geocode.us that will take an address as a parameter (stage 5). The address attribute of each of the features extracted from the http://local.yahoo.com site will be submitted to this site for processing. The resulting page will also be wrapped to extract the latitude and longitude coordinates for each address (stage 6). The spatial footprint will then be associated with the address to complete the name, type, and footprint data required for a basic gazetteer entry (stage 7).

As shown in the previous example, the ease with which these Web-extraction tools can be employed is leading to a dramatic increase in data sources for gazetteer creation and geospatial analysis. Applying these tools to the geographic domain is especially fruitful in that information typically not considered valid geographic data (e.g., phone-book data from the site http://local.yahoo.com) can be converted into data that can be mapped and interpreted geographically. As stated, the temporal footprints for the information along each of the axes may only be safely assumed "current" at each execution, but iteratively re-executing them can produce a detailed lineage of information about a feature as it exists across time periods (e.g., as tenants or uses of structures change). The ability to achieve this type of fine-scale tracking of features across time periods using spatiotemporal gazetteers has been successfully reported in research before (e.g., Agouris et al. 2000), but the advances required in the tools necessary to easily, reliably, and repeatedly generate them have until now, been missing from the literature. The types of sources described thus far provide

"spotty" temporal informa
additional data can safely
that they provide as "curre

Changing Roles for the

This progress is enabling
environments representing
was at one or more specifi
graphic interest of the pa
data sources can be wrapp
gration frameworks to pro
of the gazetteer is the mos
and tag geographic referer
entity recognition (NER)
sive list of place names for
data (i.e., alternative name
gazetteer with sparse name
Maynard et al. 2004).

Depending on the geo
ferent types of gazetteers
applications, such as ident
York Times world news s
such as the *CIA World Fa*
sufficient. In contrast, for
reports from the police blo
gazetteer covering the loc
gazetteer created using the
presently available for tha
"Robbery at Starbucks on
in a gazetteer format woul
to use a gazetteer created
gazetteer example, the ph
associated with a geospati

With this in mind, we
are particularly useful in
serve these different appli
tectures have long been u
large numbers of heteroge
kar et al. 2005). These to
sources to use for a partic
tion plan to integrate the d
gazetteer data sources and
mediator framework, one
tion capabilities developed
et al. 1996, Ambite and K

"spotty" temporal information, and further research is required to determine what additional data can safely be assumed or derived from them beyond the information that they provide as "current."

Changing Roles for the Gazetteer

This progress is enabling the rapid construction of detailed geospatial models of environments representing temporal "snapshots" (Gadia 1998) of the world as it was at one or more specific times. Depending on the geographic, temporal, or typographic interest of the party creating the gazetteer, any number of heterogeneous data sources can be wrapped and linked together using one of a number of data-integration frameworks to produce user-centric gazetteers focusing on whichever axis of the gazetteer is the most important. For example, the analysis of text to recognize and tag geographic references using natural language processing (NLP) and named entity recognition (NER) requires the use of a gazetteer that can provide an extensive list of place names for reference. For such a purpose, a gazetteer with rich name data (i.e., alternative names and spellings for places) will be a better choice than a gazetteer with sparse name data (i.e., only one name per place) (Mikheev et al. 1999, Maynard et al. 2004).

Depending on the geographic granularity of the features to be identified, different types of gazetteers can and should be created and employed. For global-scale applications, such as identifying references to countries and major cities in the *New York Times* world news section, a low-resolution gazetteer created from sources such as the *CIA World Factbook* (U.S. Central Intelligence Agency 2007) might be sufficient. In contrast, for the task of identifying geographic references in incident reports from the police blotter section of a small-town newspaper, a high-resolution gazetteer covering the local area would be required. It would be impossible for a gazetteer created using the *CIA World Factbook* (or even any well-known gazetteer presently available for that matter) to identify the geographic footprint for the text "Robbery at Starbucks on Main St," although having this more detailed information in a gazetteer format would be valuable for many purposes. In contrast, if one were to use a gazetteer created from the http://local.yahoo.com Web site in our restaurant gazetteer example, the phrase "Starbucks on Main St" could be deconstructed and associated with a geospatial footprint.

With this in mind, we will turn our attention to geospatial mediators, tools that are particularly useful in enabling the construction of user-centric gazetteers that serve these different applications. The traditional (non-geospatial) mediator architectures have long been used as a basis for information-integration projects using large numbers of heterogeneous data sources to perform complex operations (Thakkar et al. 2005). These tools excel at automatically determining the correct data sources to use for a particular problem, as well as automatically deriving an execution plan to integrate the data to achieve the desired goals. By modeling the available gazetteer data sources and the operations that take place upon those data within a mediator framework, one can leverage the heavily researched planning and execution capabilities developed in artificial intelligence for other purposes (e.g., Adali et al. 1996, Ambite and Knoblock 1997). To begin the construction of a mediator,

one describes the available information sources in terms of their input parameters (known as binding parameters), output parameters, and the functions within the system that they perform (known as domain rules) along with any constraints on their scope. This set of descriptions defines the source relations for the mediator system. The binding parameters ensure that a source cannot be queried unless all required data that is needed to proceed is available, and the domain rules represent the high-level operations performed in the mediator and are used during query reformulation. The scope constraints are used to specify minimum thresholds for the sources and operations in terms of their suitability for the task(s) at hand.

Geospatial mediators (Gupta et al. 1999, Shimada and Fukui 1999) extend the functionality of traditional mediators by embedding knowledge about geography and geospatial concepts and relationships that can be used when working with geographic data in particular. For each data source and operation, a series of both spatial and nonspatial attributes can be associated and exploited as the mediator reasons about how it should proceed to solve a particular query (Raman and Hellerstein 2002). For instance, one can associate spatial attributes such as a geographic extent (e.g., a bounding box) representing the geographic coverage for which it is valid, spatial resolution, or spatial accuracy metrics (e.g., horizontal or vertical positional accuracy), and when given a choice between multiple options, the mediator can choose the data source that provides information about the area of interest (within the geographic extent) at the highest resolution and/or level of accuracy available.

Alternatively, instead of always choosing the highest-quality data sources, the mediator can be instructed to optimize one particular characteristic over another. This might be a suitable option in the case when one wishes to optimize for response time, such as during a disaster when obtaining mostly correct information quickly is more important than obtaining entirely correct information slowly. Here, if a highly accurate source's response time is lengthy, but a less, yet sufficiently, accurate source is quicker, the mediator can reason the correct execution path and choice of sources by the simple association of the response time attribute for the data sources in its domain model.

This ability gets at the heart of the process of creating a user-centric gazetteer for specific purposes. Generalizing and embedding the gazetteer-creation process into a geospatial-mediator framework, as displayed in Figure 4.3, creates the ability to tailor gazetteers to fit particular needs.

Figure 4.3 illustrates how the gazetteer-creation process can be generalized into its three component parts, each responsible for generating one of the three required axes, representing the four domain rules of the system. The first, *GetGazetteer*(), represents the concept of creating the entire gazetteer. The remaining three, *GetNames*(), *GetTypes*(), and *GetFootprints*(), are each responsible for generating the names, types, and footprints for entries, respectively. Each of these three are themselves generalized (conceptual) functions, which in turn are further broken down into a series of composite functions $f_1()$ through $f_n()$ (that can themselves be conceptual and composed of subfunctions) that work together to achieve the desired output for that axis of the gazetteer. In this manner, the designer of gazetteer-creation processes can embed the particular requirements of their application directly into the framework for the creation of the attributes for each of the gazetteer axes.

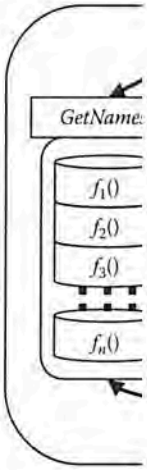


FIGURE 4.3 Generalizatio

This type of generalizatio
for a mediator framework
views over the data-sourc
this with an example. In F
the restaurant gazetteer ex
types produced be subtype
Names() and *GetTypes*() f
simplistic example, the *Ge*
three subfunctions. Param
must be satisfied before exe
dollar signs are return valu

GetGazetteer
GetTypes ($lo
GetNames ($l
GetFootprints

FIGURE 4.4 Domain rules

YahooTypes
GetTypes($l

YahooName
typedAndNa
GetNames ($

GeocoderUS

FIGURE 4.5 Source descrip

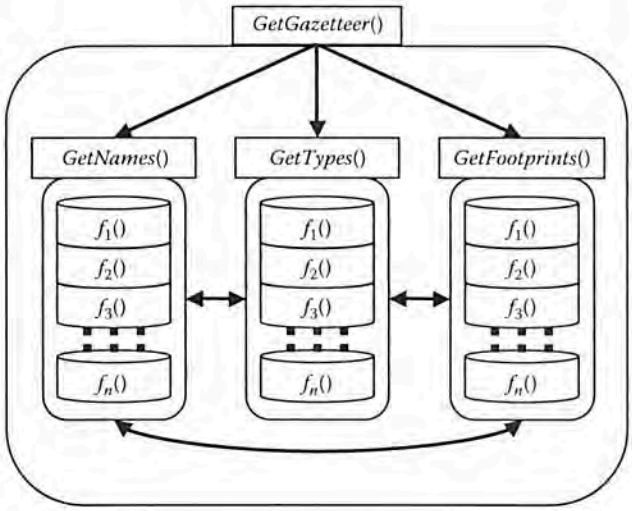


FIGURE 4.3 Generalization of gazetteer-creation process.

This type of generalization of conceptual operations just described is aptly suited for a mediator framework where the operations (domain rules) are modeled as views over the data-source models (source relations). Figures 4.4 and 4.5 ground this with an example. In Figure 4.4, the domain rules for a generalized version of the restaurant gazetteer example are provided. Here, the restriction that the feature types produced be subtypes of restaurants has been removed by decoupling the *GetNames*() and *GetTypes*() functions from the http://local.yahoo.com source. In this simplistic example, the *GetGazetteer*() function is described as a conjunction of the three subfunctions. Parameters with dollar signs are the binding parameters that must be satisfied before execution of that function can begin, and parameters without dollar signs are return values. Thus, we see that in this example the user would first

```
GetGazetteer ($location, $type, features ):–
GetTypes ($location, $type, subTypes)^
GetNames ($location, $subTypes, typedAndNamedfeatures)^
GetFootprints ($typedAndNamedfeatures, features)
```

FIGURE 4.4 Domain rules for gazetteer generation.

```
YahooTypes ($location, $type, subTypes):–
GetTypes($location, $type, subTypes)

YahooNames($location, $subTypes,
typedAndNamedfeatures):–
GetNames ($location, $subTypes, typedAndNamedfeatures)

GeocoderUS($typedAndNamedfeatures, gazetteerFeatures):–
```

FIGURE 4.5 Source description for gazetteer generation.

enter a type parameter to get the system started, which would then cause the *GetTypes*() function to execute and gather the subtypes from the http://local.yahoo.com site (that would require the creation of another separate agent that simply extracts subtypes given a supertype). Upon the generation of the list of feature types from this function, the *GetNames*() function will query for names of features with those types within the location. Finally, once the names have been associated and the features returned, the *GetFootprints*() function will associate spatial footprints with the named features to complete the gazetteer entries. Figure 4.5 displays the source relations, identifying, for each source, which domain rules it satisfies. In satisfying a query, the mediator will generate a plan that combines the available source descriptions in the proper order to fulfill the constraints of the domain relations, producing the appropriate output.

Where Are We Headed?

The restaurant gazetteer example and its generalization and implementation into a mediator plan are just samples of how current extraction and integration tools can be used to automate the discovery, extraction, and integration of geographic-data sources for the production of gazetteers. What this means for the greater research community is that vast quantities of information describing the world, geographically speaking, are at their fingertips, ready to be harvested and exploited for any number and variety of research activities. By iteratively running these newly emerging gazetteer generation tools, intelligence analysts and decision makers can use gazetteers in conjunction with other reasoning systems and tools to trace the movements of people, vehicles, goods, and services.

The automated frameworks being created for these tasks generate immediate results, extracting information from the available data sources, at the instant they are queried, which reflects the view of the world of the source. Storing the previous results of iterative runs enables a full history to be compiled of an area and how it has changed over time, thereby providing valuable information for those interested in the changing character of the geographic domain. For instance, if the workflow presented in Figure 4.1 (the simple restaurant gazetteer example) was scheduled to run every day, the results could be used to track the arrival, departure, and movement of restaurants within an area of interest. This type of information is highly descriptive of the dynamic urban commercial landscape and would be useful for those interested in data describing dynamic processes, such as economic growth patterns (as more restaurants open or close), population movements (as different classes of restaurants catering to different ethnic communities come and go), and land use (as built structures transition from one use class to another).

With a view to increasing the resolution and accuracy of the gazetteer even further, the "Wiki" concept has recently been applied to the gazetteer domain. Several online sites (e.g., http://geonames.org; GeoNames.org 2007) help individuals knowledgeable about a particular geographic area to share their knowledge in a flexible and unofficial manner by simply adding, deleting, and correcting feature entries through an Internet form. The people entering information about local areas (one would hope) are the people who actually live in and are knowledgeable about that

area. This method of gazett
level of local-scale accurac
are (currently) no restrictio
as famously shown through
nal Wikipedia.com site (Se
information about geograp
able. Although the richness
information must be weighe
able elsewhere. In addition,
entries at will, the result wil
time. These multiple views
to their usage in application
These include topics such a
the descriptive attributes as

Possible directions for
ated through new gazetteer-
checks such as secondary so
ing (Malkhi et al. 1997), a
et al. 2004). However large
the validity of these types o
makes them valuable for thei
be assumed to be legitimate.

The preceding subsectio
of the gazetteer both in ter
how the emerging data sou
in new, previously unimagin
mentally linked to the impro
geocoder. In many instance
geographic features in the g
and the strengths and weakn
on the resulting gazetteer (G
exploration of this topic.

GROUNDING THE CH

Today, most gazetteers con
Topological relations betwee
ment cannot be derived fro
administrative hierarchies (e.
country), but these same rela
prints. Large vector features,
their mouths, an enormous g
footprints. When large area f
as single points, the represent
ings have recently begun to
tools. As we will see, the im

area. This method of gazetteer-feature creation and refinement adds an entirely new level of local-scale accuracy and an entirely new set of potential problems. There are (currently) no restrictions about who can enter information about features, so, as famously shown through scandals involving the validity of the data on the original Wikipedia.com site (Seigenthaler 2005), the trust that one should place in the information about geographic features contained in a Wiki gazetteer is questionable. Although the richness of local knowledge is a plus, the unedited nature of the information must be weighed against the official and authorized information available elsewhere. In addition, when the general public is allowed to create gazetteer entries at will, the result will be many different views of a specific place at a specific time. These multiple views will generate many issues that must be addressed prior to their usage in applications that have particular accuracy or reliability constraints. These include topics such as the certainty (accuracy) and vagueness (resolution) of the descriptive attributes as well as confidence in the source.

Possible directions for addressing these and the multitude of other issues created through new gazetteer-generation techniques include applying data-reliability checks such as secondary source validation (Michalowski et al. 2005), quorum voting (Malkhi et al. 1997), and the inclusion of authoritative trusted sources (Gertz et al. 2004). However large the difficulties introduced, it is easy to imagine that the validity of these types of resources might improve over time, and this outcome makes them valuable for their accuracy and resolution as long as the data entered can be assumed to be legitimate.

The preceding subsections have attempted to describe the concept and structure of the gazetteer both in terms of its fundamental content and uses, and in terms of how the emerging data sources and methods leading to its creation are driving it in new, previously unimaginable directions. One key geospatial tool that is fundamentally linked to the improvements in gazetteer completeness and accuracy is the geocoder. In many instances, the geocoder is the tool used to create footprints for geographic features in the gazetteer (as in our prior gazetteer-generation examples), and the strengths and weaknesses of the geocoder used will have numerous effects on the resulting gazetteer (Goldberg et al. 2007). The next section offers an in-depth exploration of this topic.

GROUNDING THE CHANGING WORLD — THE GEOCODER

Today, most gazetteers contain single-point representations for their footprints. Topological relations between features are limited; for example, overlap and containment cannot be derived from points. Most gazetteers include explicit statements of administrative hierarchies (e.g., a city is part of a county is part of a state is part of a country), but these same relationships cannot be derived from the simple point footprints. Large vector features, such as rivers, are often represented by a single point at their mouths, an enormous generalization that greatly reduces the usefulness of the footprints. When large area features such as major cities or countries are represented as single points, the representation is useless for many applications. These shortcomings have recently begun to be addressed through more sophisticated geocoding tools. As we will see, the improvements made to the underlying methodology, the

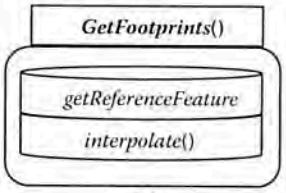
inclusion of new types of data sources, and the expansion (or perhaps better termed, relaxation) of the overarching concept of geocoding all play a role in improving the usefulness and accuracy of geocoding processes.

Creating a Truer Representation of the Dynamic World

Improvements made to geocoder tools can most easily be seen through both the accuracy and representation of their outputs. When considering (for simplicity) only point output, more accurate means that the location of the point produced will be closer to where it should be on the ground. Improving the output of geocoders requires rethinking what its output should be, in order to meet user requirements. Although single points are sufficient for some purposes (e.g., orientation of a map view) other analyses require footprints that bound an area or trace a linear feature. Some projects require three-dimensional (3D — length, width, height) or four-dimensional (4D — length, width, height, time) representations. In the case of geocoding an address, this could correspond to returning the 3D structure of the building as it was at a particular time (making it 4D), or the 3D structure of the hills and valleys at a particular time (4D). Other considerations for improving geocoder output are using more accurate reference sources (Ratcliffe 2001, Cayo and Talbot 2003) and eliminating assumptions in the geocoding process (Bakshi et al. 2004, Goldberg et al. 2007).

Many of the same advancements leading to the production of more complete and accurate gazetteers (e.g., availability of data sources, employment of mediator technologies) are driving advances in geocoding technologies on both fronts: accuracy and representation (e.g., Bakshi et al. 2004). This, in turn, is leading to the production of gazetteers whose features are more representative of the real world, with topological relations between features intact. Attacking the very assumptions that have traditionally caused the spatial inaccuracies in the geocoding process, new tools supporting new algorithms for deriving a geographic point from a textual piece of information are vastly improving the results of the geocoding process. Both the address range (Ratcliffe 2001) and parcel homogeneity (Dearwent et al. 2001) assumptions that plague linear interpolation-based geocoding solutions can be avoided by using the same information extraction tools mentioned earlier. Figures 4.6 and 4.7 show how the *GetFootprints()* domain rule can be expanded with new data sources and operations resulting in both highly accurate point representations and alternative polygon footprints depending on user needs. Figure 4.6 conceptually depicts multiple versions of the same function. Figure 4.7 shows how these could be implemented in the mediator system.

By wrapping nontraditional, highly accurate, local-scale information sources that can validate the existence and location of addresses along a street segment (such as local tax assessor Web sites), geocoding tools can generate geographic coordinates for addresses based on the actual number of parcels on the street, rather than by using the less-accurate address ranges associated with street vector sources. Sometimes, the dimensions of parcel lots along a street are available and, with this knowledge in hand, geocoding algorithms no longer need to assume that all parcels along a street segment are the same size (the parcel homogeneity assumption). Instead, the



(a)

FIGURE 4.6 Alternative expa

algorithm can reason over th
cels on the entire block (all f
mine the most probable layo
corners (the so-called corne
smaller domain of error, ulti
geocoded location (Bakshi e
situation is presented. In the

```
GetFootprints($address, $bo
GetFootprintsLinear ($ addr
GetFootprintsLinearAdvanc
GetFootprintsExtract ($ add
```

```
GetFootprintsLinear ($ addr
GetReferenceFeature($addre
Interpolate($address, $ref, fe
boundingBox='US'^
minAccuracy='0'
```

```
GetFootprintsLinearAdvanv
GetReferenceFeature($addre
GetSegmentsOnBlock($ref, b
GetNumberOfHouses($ref, n
GetDimensions($ref, dim)^
Interpolate($address, $ref, $b
boundingBox='LA County'^
minAccuracy='5'
```

```
GetFootprintsExtract($ addr
GetParcelMap($address, map
ExtractPolygon(map, feature)
boundingBox='Santa Monica
minAccuracy='9'
```

FIGURE 4.7 Expanded GetFo

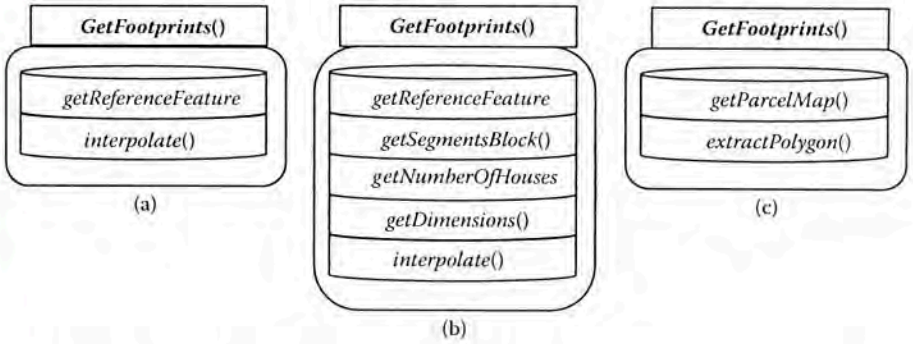


FIGURE 4.6 Alternative expansions of GetFootprints().

algorithm can reason over the length of the street segments and the sizes of the parcels on the entire block (all four sides in the case of a normal square block) to determine the most probable layout for the parcels, including which parcels are on which corners (the so-called corner problem). From this, interpolation can occur within a smaller domain of error, ultimately leading to far greater accuracy for the resulting geocoded location (Bakshi et al. 2004). In the second depiction (Figure 4.7b), this situation is presented. In the third depiction (Figure 4.7c), we see that we can just as

```
GetFootprints($address, $boundingBox:'US', $minAccuracy:'0', feature):-
GetFootprintsLinear ($ address, feature) V
GetFootprintsLinearAdvanced ($ address, feature) V
GetFootprintsExtract ($ address, feature) V
```

```
GetFootprintsLinear ($ address, $boundingBox, feature):-
GetReferenceFeature($address, ref)^
Interpolate($address, $ref, feature)^
boundingBox='US'^
minAccuracy='0'
```

```
GetFootprintsLinearAdvanved ($ address, $boundingBox, feature):-
GetReferenceFeature($address, ref)^
GetSegmentsOnBlock($ref, block)^
GetNumberOfHouses($ref, num)^
GetDimensions($ref, dim)^
Interpolate($address, $ref, $block, $num, $dim, feature)^
boundingBox='LA County'^
minAccuracy='5'
```

```
GetFootprintsExtract($ address, $boundingBox, feature):-
GetParcelMap($address, map)^
ExtractPolygon(map, feature)^
boundingBox='Santa Monica'^
minAccuracy='9'
```

FIGURE 4.7 Expanded GetFootprints() domain models.

```
Tiger($address, ref):–
GetRerferenceFeature($ address)^
GetSegmentsOnBlock($ref, block)^
boundingBox = 'US'
```

```
InterpolationEngine($ address, $ref, feature):–
Interpolate($address, $ref, feature)
```

```
Assessor($ address, $ref, num, dim):–
GetNumberOfHouses($ref, num)^
GetDimensions($ref, dim)^
boundingBox = 'LA County'
```

```
CityMap($ address, map):–
GetParcelMap($address, map)^
boundingBox='Santa Monica'
```

```
ExtractionEngine($ map, feature):–
ExtractPolygon($map, feature)
```

FIGURE 4.8 Expanded GetFootprints() source descriptions.

easily substitute a different composite function that simply extracts the parcel polygon from an online raster map.

The additional (optional) parameters of $boundingBox and $minAccuracy have also been added in Figure 4.7, constraining the execution of each function prototype (i.e., the scope constraints), with default values assumed if absent. The $minAccuracy parameter relates to a relative level of accuracy associated with each geocoding method, 0 being the least accurate and 9 being the most. The $boundingBox parameter relates to the source descriptions defined in Figure 4.8 and defines the area for which each source is applicable. For clarity, textual strings (i.e., "LA County") are used in place of the actual geographic footprints that would be used in practice. When a geocoding query is presented to the geospatial mediator, the system will use the domain rules in Figure 4.7 and the source relations in Figure 4.8 to reason which sources and methods are appropriate to query and execute based on the required level of accuracy and availability of data. In this way, a geospatial mediator-based geocoding system is easily expandable as new data sources become available and new geocoding techniques are developed. The system will automatically determine the best methods and data sources to use based on the attributes associated with each. Figure 4.9 depicts extracting the parcel boundaries from a raster Web-map assessor Web site to return a polygon as the result of the most accurate geocoding method.

It is equally important to note that these additional data sources can be queried every time a geocoding operation is to be performed. Being national-scale and released every 10 years, the Census's TIGER Line files (a commonly used and freely available street-vector data source), like many other data sources used as a reference file for geocoding, represent static views of the world. They are temporally fixed to the time period when they were created, with deteriorating quality as time passes. This unavoidable degradation over time ultimately leads to geocoding results that

FIGURE 4.9 Parcel extracti

become more and more ina
along with the data produc
geocoding tools using mul
the example in Figures 4.6
sent the current state of the

Beyond Traditional Geo

As mentioned in the previ
levels of spatial represent
in the gazetteers that they
tioned police-blotter exam
spatial analysis carried out
for a particular area (Cha
were geared to more of an
responding to a report nee
rials being released into t
representation of a buildin

xtracts the parcel poly-

nd $minAccuracy have
:ach function prototype
ıbsent. The $minAccu-
ed with each geocoding
$boundingBox param-
ınd defines the area for
(i.e., "LA County") are
ld be used in practice.
ıtor, the system will use
ure 4.8 to reason which
based on the required
ospatial mediator-based
; become available and
ıtomatically determine
es associated with each.
ıster Web-map assessor
e geocoding method.
ta sources can be que-
eing national-scale and
nmonly used and freely
rces used as a reference
are temporally fixed to
quality as time passes.
geocoding results that

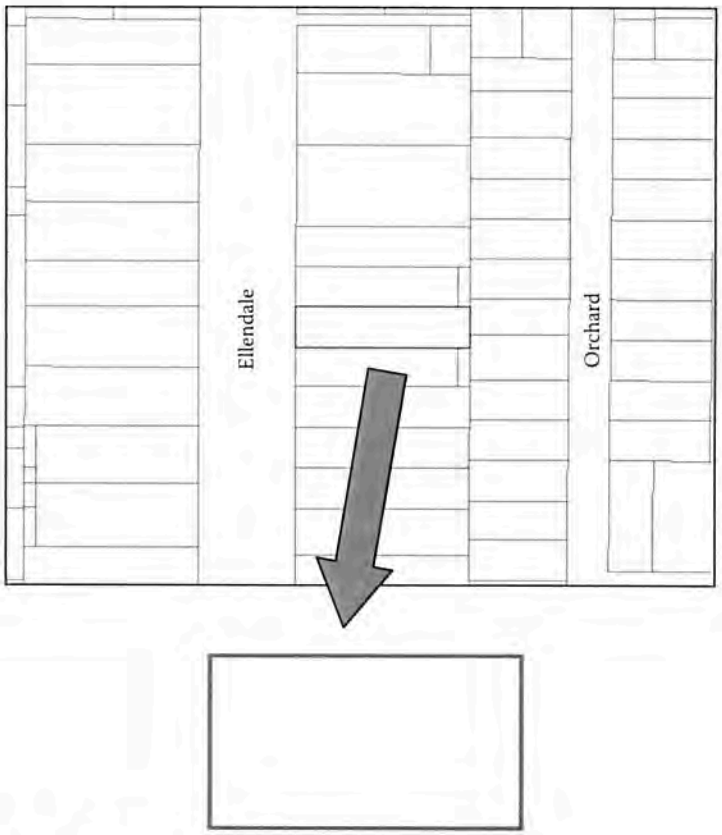


FIGURE 4.9 Parcel extraction.

become more and more inaccurate (Bonner et al. 2003), a fact that is seldom reported along with the data produced (Rushton et al. 2006). In contrast, the newly emerging geocoding tools using multiple sources for feature creation and validation (such as the example in Figures 4.6 and 4.7 based on geospatial mediator technologies) represent the current state of the same geocoded location.

Beyond Traditional Geocoding

As mentioned in the previous sections, different applications will require varying levels of spatial representations and resolutions for the geographic data contained in the gazetteers that they use for their data sources. For instance, in the aforementioned police-blotter example, a simple point representation may be sufficient for the spatial analysis carried out by police investigators into patterns of crime occurrence for a particular area (Chainey and Ratcliffe 2005). In contrast, if the application were geared to more of an emergency-management situation and the police officers responding to a report needed to know information about potential hazardous materials being released into the environment for evacuation purposes, a single-point representation of a building would not suffice. In this case, the information required

would be the two-dimensional (2D) polygon representation or even the three-dimensional (3D) model of the entire building at some predefined level of accuracy. This more detailed information about the location would enable emergency management personnel to predict the dispersion of toxins with greater accuracy, enabling them to direct evacuations of people in the most dangerous locations first.

The availability of gazetteers containing subparcel, 3D, and indoor features (derived from geocoding tools) was unimaginable just a few years ago. The proliferation of new types of online data such as publicly available raster maps of assessor parcels and 3D building models from data sources like Google Earth are facilitating entirely new tiers of geocoding capabilities (e.g., Hutchinson and Veenendall 2005a, b; Lee 2004). Once limited by the availability of data, the geocoding tools are growing rapidly in terms of what can be geocoded and to what level of accuracy. As one example, through a combination of computer vision and information extraction techniques, researchers are beginning to extract feature boundaries reliably from online maps available freely to the public via the Internet. Separating the composite layers into individual geographic features, the process of geocoding is moving beyond simple and traditional linear interpolation to actual geographic-feature extraction with true geographic boundaries intact (e.g., Chiang et al. 2005). This represents a significant advance in both the accuracy and representation of the geocoded features produced, providing an unprecedented feature-by-feature view of the changing world. As features change over time, each of these representations can be captured, providing accurate data from which specific events and other complex space-time relationships can be derived.

Additionally, the types of information that can be geocoded are rapidly increasing. Long the bane of geocoding practitioners, for example, sub-parcel geocoding is quickly becoming a reality (Hutchinson and Veenendall 2005a, b; Lee 2004). The importance of this to researchers of all stripes cannot be understated. In the health-research community alone, accurate representations of a person's domicile, down to the apartment level, will serve to promote entire new lines of research that were heretofore unavailable due to the availability of geocoded data at too-low resolution (e.g., down to "somewhere" in the parcel of the apartment complex's footprint). Environmental exposure models sensitive down to meters or tens of meters can now be reliably created, applied, and validated with greater confidence because of the more accurate data being produced (e.g., Ward et al. 2005).

Perhaps even more important, the geocoding of relative spatial locations is proving feasible. Locational descriptions such as "half a mile north of Tutor Hall," being both sub-parcel (Tutor Hall on USC Campus) and relative (half mile), have never been reliably geocoded by any commercial geocoding method. Prototypical research platforms have investigated the possibility, but production systems are highly specialized for a particular area, and operational systems can be found for only a handful of places. Combining fundamental research from the fields of machine learning, artificial intelligence, and GIS, emerging research is providing the framework for this type of tool to become a reality (cf. Hutchinson and Veenendall 2005a, b). When it comes to fruition, this line of research will provide immeasurable utility to the most basic of geographical concepts, explaining the location of something in simple terms. When a call comes into a 911 center, unless a person is at his or her own home

or work, he or she will pro
she needs assistance, much
easily in terms of relative lo
accurately is of utmost imp
credible intelligence reports

DYNAMIC INTEGRATI

The dynamic modeling of t
oding services provide is j
being useful for data genera
the integration of disparate,
pretation of the changing wo
tools brings different streng
therefore complement one a

Enabling Integration wit

The primary strength of the
the identification of a geog
axes (name, type, or footpri
assume a central role in dyn
possesses the ability to link
simple fact that any inform
to at least one of the three
dynamically changing data
one or more political activi
all of the stream gauging st
source that are downstream

We can see through the
the full spectrum of possible
features (e.g., spatially, tem
as the basis for reasoning wi
ing a data set with the gaze
gazetteer), it becomes imme
well. A set of place names c
by intersecting along the to
have names and types associ
While there may be cases w
to the wrong scale or resolut
name "United States" assoc
world countries), this same
to the correct situation using

Likewise, advancement
ate access to unprecedente
often-cited quotation used

or even the three-dimen-
ed level of accuracy. This
emergency management
ccuracy, enabling them to
ns first.
3D, and indoor features
ew years ago. The prolif-
le raster maps of assessor
ogle Earth are facilitating
on and Veenendall 2005a,
geocoding tools are grow-
level of accuracy. As one
formation extraction tech-
aries reliably from online
rating the composite lay-
coding is moving beyond
graphic-feature extraction
l. 2005). This represents
tion of the geocoded fea-
ture view of the changing
entations can be captured,
other complex space-time

coded are rapidly increas-
e, sub-parcel geocoding is
2005a, b; Lee 2004). The
nderstated. In the health-
person's domicile, down
ines of research that were
data at too-low resolution
complex's footprint). Envi-
ens of meters can now be
dence because of the more

e spatial locations is prov-
north of Tutor Hall," being
ve (half mile), have never
hod. Prototypical research
n systems are highly spe-
be found for only a hand-
ields of machine learning,
viding the framework for
eenendall 2005a, b). When
nmeasurable utility to the
on of something in simple
n is at his or her own home

or work, he or she will probably not be able to give an exact address where he or she needs assistance, much less the exact geographical location. People speak most easily in terms of relative locations and as such, a tool to geocode them quickly and accurately is of utmost importance. Responding to unexpected events and building credible intelligence reports in a dynamically changing world will necessitate it.

DYNAMIC INTEGRATION

The dynamic modeling of the environment that newly emerging gazetteer and geocoding services provide is just one outcome that these tools afford. In addition to being useful for data generation and representation, these tools are also facilitating the integration of disparate, dynamic data sources for use in the analysis and interpretation of the changing world they represent (cf. Shahabi et al. 2006). Each of these tools brings different strengths to the problem of dynamic data integration and they therefore complement one another.

Enabling Integration with Gazetteers

The primary strength of the classical gazetteer conceptual structure is that it enables the identification of a geographic feature from information along any of its three axes (name, type, or footprint) (Hill 2000). This structure enables the gazetteer to assume a central role in dynamic data integration. First and foremost, the gazetteer possesses the ability to link a variety of different geographic references due to the simple fact that any information about a geographic feature of interest must refer to at least one of the three axes. For example, the gazetteer could be used to link dynamically changing data about political stability with the probable locations of one or more political activists. Alternatively, a gazetteer might be used to identify all of the stream gauging stations based on the feature types in a hydrological data source that are downstream of a dam that is likely to fail.

We can see through the previous examples that, because the gazetteer cuts across the full spectrum of possible information that can be used for describing geographic features (e.g., spatially, temporally, toponymically, typographically), it can be used as the basis for reasoning with a wide variety of geospatial data sources. By integrating a data set with the gazetteer (e.g., intersecting along the appropriate axis of the gazetteer), it becomes immediately and automatically grounded to the other axes as well. A set of place names can be associated with spatial footprints and types simply by intersecting along the toponymic axis. Likewise, a set of spatial footprints can have names and types associated with them by integrating the two sets of geometries. While there may be cases where the automatic association of information might be to the wrong scale or resolution to be useful (e.g., every parcel in a city will have the name "United States" associated with it if intersected spatially with a gazetteer of world countries), this same fundamental ability can prove invaluable when applied to the correct situation using the correct data sources.

Likewise, advancements made to geocoding services are enabling immediate access to unprecedented amounts of information based on address data. An often-cited quotation used primarily to validate the time and money being spent

on geocoding research projects is that "80–90% of [U.S.] government data is geographic data" (Federal Geographic Data Committee 2006), with the majority of that consisting of raw address data or geocoded address data (Croner 2003). From this standpoint alone, we could rightfully conclude that address data could be considered among the most (if not the most) ubiquitous form of geospatial data in existence. Like never before, geographic references that lead to any form of addressable data can now be integrated immediately with other spatial data sets, through the contents of a gazetteer database and/or the use of geocoding technologies to convert a textual address (nongeographic data) into some form of geospatial data (e.g., a point, line, or feature).

CONCLUSIONS

Taken together, the newly emerging gazetteer and geocoding tools being developed provide a powerful combination with which to both describe and work with dynamic data in a changing world. Fundamental research advances in both of these interrelated fields are leading the way in both information generation and information integration, each of which are of particular concern to intelligence analysts and decision makers as they try to explain the trends and events that characterize our continually evolving world.

REFERENCES

Adali, S., K. S. Candan, Y. Papakonstantinou, and V. S. Subrahmanian. 1996. VS. Query caching and optimization in distributed mediator systems. In *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, 137–46. New York: ACM Press.

Agarwal, P. 2004. Contested Nature of Place: Knowledge Mapping for Resolving Ontological Distinctions Between Geographical Concepts. In *GIScience 2004*, ed. M. J. Egenhofer, C. Freksa, and H. J. Miller, 1–21. Lecture Notes in Computer Science 3234. Berlin: Springer-Verlag.

Agouris, P., K. Beard, G. Mountrakis, and A. Stefanidis. 2000. Capturing and Modeling Geographic Object Change: A SpatioTemporal Gazetteer Framework. *Photogrammetric Engineering and Remote Sensing* 66(10):1224–50.

Alexandria Digital Library. 2007a. http://www.alexandria.ucsb.edu/. Last accessed 3/25/2007.

Alexandria Digital Library. 2007b. *Guide to the ADL Gazetteer Content Standard*, Version 3.2. http://www.alexandria.ucsb.edu/gazetteer/ContentStandard/version3.2/GCS3.2-guide.htm. Last accessed 3/25/2007.

Ambite, J. L. and C. A. Knoblock. 1997. Planning by Rewriting: Efficiently Generating High-Quality Plans. In *Proceedings of the 14th National Conference on Artificial Intelligence*. Providence, RI: AAAI Press / MIT Press. 706–13.

Axelrod, A. 2003. On Building a High Performance Gazetteer Database. In A. Kornai and B. Sundheim (eds.): HLT/NAACL '03: *Proceedings of Workshop on the Analysis of Geographic References held at Joint Conference for Human Language Technology and Annual Meeting of the North American Chapter of the Association for Computational Linguistics*. Boston, MA: Association for Computational Linguistics. 63–68.

Bakshi, R., C. A. Knoblock, a
Geocode Addresses. In
Proceedings of the 12th
Information Systems. W

Barish, G. and C. A. Knoblock
tem for Software Agents

Bonner, M. R., D. Han, J. Nie
tional Accuracy of Geo
14(4). 408–11.

Buckland, M. and L. Lancast
Cultural Atlas Initiativ
buckland/05buckland.ht

Cayo, M. R. and T. O. Talbot.
addresses. *International*

Chainey, S. and J. Ratcliffe. 2

Chiang, Y-Y., C. A. Knoblock
sections from raster ma
in Geographic Informa
267–76.

Croner, C. M. 2003. Public He
57–82.

Dearwent, S. M., R. R. Jacobs
encing public health dat
ogy 11(4):329–34.

Doerr, M. 2001. Semantic Pro
1(8). 2001–2003.

Doerr, M. and M. Papagelis.
matching. Unpublished,
http://www.ics.forth.gr/i

Dugandzic, R., L. Dodds, D.
low level exposures to a
cohort study. *Environme*

Federal Geograhpic Data Co
tion Systems: How GIS
post-September 11th An
papers/homeland-securi

Frew, J., M. Freeston, N. Freit
Q. Zheng. 1998. The Al
Technology for Digital I
(ECDL). 61–73.

Frew, J., M. Freeston, N. Freit
Q. Zheng. 2000. The A
on Digital Libraries 2(4

Frank, A. U. 2001. Tiers of ont
systems. *International J*

Gadia, S. K. 1998. A homoge
bases. *ACM Transactio*
418–48.

GeoNames.org. 2007. http://w

Gertz, M., M. T. Ozsu, G. Sa
Data quality on the Web

Bakshi, R., C. A. Knoblock, and S. Thakkar. 2004. Exploiting Online Sources to Accurately Geocode Addresses. In D. Pfoser, I.F. Cruz, and M. Ronthaler (eds.): *ACM-GIS '04: Proceedings of the 12th ACM International Symposium on Advances in Geographic Information Systems*. Washington, D.C.: ACM Press. 194–203.

Barish, G. and C. A. Knoblock. 2005. An Expressive Language and Efficient Execution System for Software Agents. *Journal of Artificial Intelligence Research* 23:625–66.

Bonner, M. R., D. Han, J. Nie, P. Rogerson, J. E. Vena, and J. L. Freudenheim. 2003. Positional Accuracy of Geocoded Addresses in Epidemiologic Research. *Epidemiology* 14(4). 408–11.

Buckland, M. and L. Lancaster. 2004. Combining Place, Time, and Topic: The Electronic Cultural Atlas Initiative. *D-Lib Magazine* 10(5). http://www.dlib.org/dlib/may04/buckland/05buckland.html.

Cayo, M. R. and T. O. Talbot. 2003. Positional error in automated geocoding of residential addresses. *International Journal of Health Geographics* 2(10).

Chainey, S. and J. Ratcliffe. 2005. *GIS and Crime Mapping*. Chichester, UK: Wiley & Sons.

Chiang, Y-Y., C. A. Knoblock, and C-C. Chen. 2005. Automatic extraction of road intersections from raster maps. In *The 13th ACM International Symposium on Advances in Geographic Information Systems* (ACM-GIS'05), Bremen, Germany, November. 267–76.

Croner, C. M. 2003. Public Health GIS and the Internet. *Annual Review of Public Health* 24: 57–82.

Dearwent, S. M., R. R. Jacobs, and J. B. Halbert. 2001. Locational uncertainty in georeferencing public health datasets. *Journal of Exposure Analysis Environmental Epidemiology* 11(4):329–34.

Doerr, M. 2001. Semantic Problems of Thesaurus Mapping. *Journal of Digital Information* 1(8). 2001–2003.

Doerr, M. and M. Papagelis. 2006. A method for estimating the precision of place name matching. Unpublished, under review to *International Journal of Digital Libraries*. http://www.ics.forth.gr/isl/publications/paperlink/ecdl.pdf. Last accessed 11/24/2006.

Dugandzic, R., L. Dodds, D. Stieb, and M. Smith-Doiron. 2006. The association between low level exposures to ambient air pollution and term low birth weight: a retrospective cohort study. *Environmental Health* 5 (3).

Federal Geograhpic Data Committee. 2006. Homeland Security and Geographic Information Systems: How GIS and mapping technology can save lives and protect property in post-September 11th America. http://www.fgdc.gov/library/whitepapers-reports/whitepapers/homeland-security-gis. Last accessed 11/24/2006.

Frew, J., M. Freeston, N. Freitas, L. L. Hill, G. Janée, K. Lovette, R. Nideffer, T. Smith, and Q. Zheng. 1998. The Alexandria Digital Library Architecture. Research and Advanced Technology for Digital Libraries. In *Proceedings of the Second European Conference* (ECDL). 61–73.

Frew, J., M. Freeston, N. Freitas, L. L. Hill, G. Janée, K. Lovette, R. Nideffer, T. Smith, and Q. Zheng. 2000. The Alexandria Digital Library Architecture. *International Journal on Digital Libraries* 2(4):259–68.

Frank, A. U. 2001. Tiers of ontology and consistency constraints in geographical information systems. *International Journal of Geographical Information Science* 15(7):667–78.

Gadia, S. K. 1998. A homogeneous relational model and query languages for temporal databases. *ACM Transactions on Database Systems* 13(4). New York, NY: ACM Press. 418–48.

GeoNames.org. 2007. http://www.geonames.org, Last accessed 3/25/2007.

Gertz, M., M. T. Ozsu, G. Saake, and K. U. Sattler. 2004. Report on the Dagstuhl Seminar: Data quality on the Web. *SIGMOD Record* 33(1). New York, NY: ACM Press. 127–32.

Goldberg, D. G., J. P. Wilson, and C. A. Knoblock. 2007. From Text to Geographic Coordinates: The Current State of Geocoding. *URISA Journal* 19(1):33–46.

Goodchild, M. 1999. The future of the gazetteer. Unpublished, presented at the Digital Gazetteer Information Exchange Workshop. http://www.alexandria.ucsb.edu/~lhill/dgie/DGIE_website/DGIE final report.htm. Last accessed 11/24/2006.

Goodchild, M. and L. L. Hill. 2007. *Summary Report: Digital Gazetteer Research & Practice Workshop*, December 7–9, 2006, Santa Barbara, CA. http://www.ncgia.ucsb.edu/projects/nga/docs/DGRP-summary-report.pdf. Last Accessed 3/25/2007.

Gupta, A., R. Marciano, I. Zaslavsky, and C. Baru. 1999. Integrating GIS and Imagery through XML-Based Information Mediation. *Integrated Spatial Databases: Digital Images and GIS*. Volume 1737 of Lecture Notes in Computer Science. Berlin: Springer. 211–34.

Hill, L. L., J. Frew, and Q. Zheng. 1999. *Geographic names: The implementation of a gazetteer in a georeferenced digital library*. Technical report.

Hill, L. L. and Zheng Q. 1999. Indirect geospatial referencing through place names in the digital library: Alexandria digital library experience with developing and implementing gazetteers. In *62nd Annual Meeting of the American Society for Information Science*. Medford, NJ: Information Today. 57–69.

Hill, L. L. 2000. Core elements of digital gazetteers: Placenames, categories, and footprints. In J. L. Borbinha and T. Baker (eds.): *ECDL '00: Research and Advanced Technology for Digital Libraries, 4th European Conference*. Volume 1923 of Lecture Notes in Computer Science. London, UK: Springer. 280–90.

Hill, L. L. 2006. *Georeferencing: The Geographic Associations of Information*. Cambridge, MA: MIT Press.

Himmelstein, M. 2005. Local Search: The Internet Is the Yellow Pages. *Computer* 38(2). Piscataway, NJ: IEEE Press. 26–34.

Hornsby, K. 2001. Temporal Zooming. *Transactions in GIS* 5(3):255–72.

Hutchinson, M. and B. Veenendall. 2005a. Towards a Framework for Intelligent Geocoding. In *SSC 2005 Spatial Intelligence, Innovation and Praxis: The National Biennial Conference of the Spatial Sciences Institute*.

Hutchinson, M. and B. Veenendall. 2005b. Towards Using Intelligence to Move from Geocoding to Geolocating. In *7th Annual GIS in Addressing Conference*. Park Ridge, IL: Urban and Regional Information Systems Association.

Knoblock, C.A., K. Lerman, S. Minton, and I. Muslea. 2001. A machine-learning approach to accurately and reliably extracting data from the Web. In *Proceedings of the IJCAI Workshop on Adaptive Text Extraction and Mining*.

Laender, A.H.F., K.A.V. Borges, J.C.P. Carvalho, C. B. Medeiros, A. S. da Silva, and C. A. Davis Jr. 2005. In Y. Manalopoulos and A. N. Papadapoulos (eds.): Integrating Web Data and Geographic Knowledge into Spatial Databases. *Spatial Databases: Technologies, Techniques and Trends*. Chapter 2:23–47.

Laender, A.H.F., B. Ribeiro-Neto, A. Silva, and J. Teixeira. 2002. A Brief Survey of Web Data Extraction Tools. *SIGMOD Record* 31(2). New York, NY: Association for Computing Machinery. 84–93.

Lam, N. S-N. and D. A. Quattrochi. 1992. On the issues of scale, resolution, and fractal analysis in the mapping sciences. *The Professional Geographer* 44(1):88–98.

Lee, J. 2004. 3D GIS for Geo-coding Human Activity in Micro-scale Urban Environments. In M. J. Egenhofer, C. Freksa, and H. J. Miller (eds.): Geographic Information Science: Third International Conference, *GIScience 2004*, 162–78. Berlin: Springer-Verlag.

Malkhi, D., M. Reiter, and R. Wright. 1997. Probabilistic quorum systems. In *PODC '97: Proceedings of the sixteenth annual ACM symposium on principles of distributed computing*. New York, NY: ACM Press. 267–73.

Maynard, D., K. Bontcheva, a
Induction of Gazetteer L
tion Conference (LREC'

Michalowski, M., S. Thakkar,
ary sources to align info
AAAI Press. 33–44.

Mikheev, A., M. Moens, and (
In H. S. Thompson and
pean Chapter of the As
Morgan Kaufmann Publ

Naumann, F. and M. Haeussle
Proceedings of the Inter

Plewe, B. 2002. The Nature o
tions in GIS 6(4):431–56

Raman, V. and J. M. Hellerst
MOD '02: *Proceedings*
agement of data. New Y

Ratcliffe, J. H. 2001. On the
cadastral and census ar
Science 15(1):473–85.

Rushton, G., M. Armstrong, J
2006. Geocoding in Ca
Medicine 30(2):S16–S2

Seigenthaler, J. 2005. A fals
com/news/opinion/edito

Shahabi, C., Y-Y. Chiang, K.
lock, S. C. Lee, U. Neum
Dec: Enabling Geospati
Conference on Multime

Shimada, S. and H. Fukui. 1
Transfer Interface in Si3
ference on Interoperati
Verlag. 265–76.

Smith, D.A. and G. S. Mann.
Sundheim (eds.): *HLT/*
graphic References hel
Annual Meeting of the
Linguistics. Boston, M

Thakkar, S., J. L. Ambite, and
plans for bioinformatics

Turmo, J., A. Ageno, and N. C
Surveys 38(4). New Yor

U.S. Board on Geographic Na
names.usgs.gov/pls/gnis

U.S. Central Intelligence Age
Central Intelligence Ag

U.S. National Geospatial-Inte
nga.mil/geonames/GNS

Van Rijsbergen, C. J. 1979. *Inf*

Wang, J. and N. Ge. 2006. Au
from digital gazetteer.
ference on Digital libra

Maynard, D., K. Bontcheva, and H. Cunningham. 2004. Automatic Language-Independent Induction of Gazetteer Lists. In *Proceedings of 4th Language Resources and Evaluation Conference* (LREC'04).

Michalowski, M., S. Thakkar, and C. A. Knoblock. 2005. Automatically utilizing secondary sources to align information across sources. *AI Magazine* 26(1). Menlo Park, CA: AAAI Press. 33–44.

Mikheev, A., M. Moens, and C. Grover. 1999. Named entity recognition without gazetteers. In H. S. Thompson and A. Lascarides (eds.): *EACL '99: 9th Conference of the European Chapter of the Association for Computational Linguistics*. San Francisco, CA: Morgan Kaufmann Publishers. 1–8.

Naumann, F. and M. Haeussler. 2002. Declarative Data Merging with Conflict Resolution. In *Proceedings of the International Conference on Information Quality* (IQ 2002).

Plewe, B. 2002. The Nature of Uncertainty in Historical Geographic Information. *Transactions in GIS* 6(4):431–56.

Raman, V. and J. M. Hellerstein. 2002. Partial results for online query processing. In SIGMOD '02: *Proceedings of the 2002 ACM SIGMOD international conference on management of data*. New York, NY: ACM Press. 275–86.

Ratcliffe, J. H. 2001. On the accuracy of TIGER-type geocoded address data in relation to cadastral and census areal units. *International Journal of Geographical Information Science* 15(1):473–85.

Rushton, G., M. Armstrong, J. Gittler, B. Greene, C. Pavlik, M. West, and D. Zimmerman. 2006. Geocoding in Cancer Research — A Review. *American Journal of Preventive Medicine* 30(2):S16–S24.

Seigenthaler, J. 2005. A false Wikipedia 'biography.' USA Today. http://www.usatoday.com/news/opinion/editorials/2005-11-29-wikipedia-edit_x.htm.

Shahabi, C., Y-Y. Chiang, K. Chung, K-C. Huang, J. Khoshgozaran-Haghighi, C. A. Knoblock, S. C. Lee, U. Neumann, R. Nevatia, A. Rihan, S. Thakkar, and S. You. 2006. GeoDec: Enabling Geospatial Decision Making. In *Proceedings of the IEEE International Conference on Multimedia & Expo* (ICME).

Shimada, S. and H. Fukui. 1999. Geospatial Mediator Functions and Container-Based Fast Transfer Interface in Si3CO Test-Bed. In *Proceedings of the Second International Conference on Interoperating Geographic Information Systems*. London, UK: Springer-Verlag. 265–76.

Smith, D.A. and G. S. Mann. 2003. Bootstrapping toponym classifiers. In A. Kornai and B. Sundheim (eds.): *HLT/NAACL '03: Proceedings of Workshop on the Analysis of Geographic References held at Joint Conference for Human Language Technology and Annual Meeting of the North American Chapter of the Association for Computational Linguistics*. Boston, MA: Association for Computational Linguistics.

Thakkar, S., J. L. Ambite, and C. A. Knoblock. 2005. Composing, optimizing, and executing plans for bioinformatics web services. *VLDB Journal* 14(3):330–53.

Turmo, J., A. Ageno, and N. Català. 2006. Adaptive information extraction. *ACM Computing Surveys* 38(4). New York, NY: ACM Press.

U.S. Board on Geographic Names. 2006. Geographic Names Information System. http://geonames.usgs.gov/pls/gnispublic. Last accessed 1/24/2006.

U.S. Central Intelligence Agency. 2007. *The World Factbook, 2007*. Washington, D.C.: U.S. Central Intelligence Agency.

U.S. National Geospatial-Intelligence Agency. 2006. GeoNet Names Server. http://gnswww.nga.mil/geonames/GNS. Last accessed 11/24/2006.

Van Rijsbergen, C. J. 1979. *Information Retrieval*. Newton, MA, USA: Butterworth-Heinemann.

Wang, J. and N. Ge. 2006. Automatic feature thesaurus enrichment: extracting generic terms from digital gazetteer. In *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*. New York, NY: ACM Press. 326–33.

Ward, M. H., J. R. Nuckols, J. Giglierano, M. R. Bonner, C. Wolter, M. Airola, W. Mix, J. S. Colt, and P. Hartge. 2005. Positional Accuracy of Two Methods of Geocoding. *Epidemiology* 16(4):542–47.

WikMapia.org. 2007. http://www.wikimapia.org, Last accessed 3/25/2007.

Yahoo! Inc. 2006. Yahoo! Local Pages. http://local.yahoo.com. Last accessed 11/24/2006.

5 Recc Indi Expc Envi Con Usin

Jaymie R

CONTENTS

"I sometimes ask myself h
relativity. The reason, I thi
lems of space and time. T
my intellectual developme
space and time only when