Geoinformatica (2007) 11:103–129
DOI 10.1007/s10707-006-0015-7

Assessing the Certainty of Locations Produced by an Address Geocoding System

Clodoveu A. Davis Jr · Frederico T. Fonseca

Received: 3 February 2006 / Revised: 6 July 2006 /
Accepted: 3 November 2006 / Published online: 13 January 2007

Abstract Addresses are the most common georeferencing resource people use to communicate to others a location within a city. Urban GIS applications that receive data directly from citizens, or from legacy information systems, need to be able to quickly and efficiently obtain a spatial location from addresses. In this paper we understand addresses in a broader perspective, in which not only the conventional elements of postal addresses are considered, but other kinds of direct or indirect references to places, such as building names, postal codes, or telephone area codes, which are also valuable as locators to urban places. This broader view on addresses allows us to work with two perspectives. First, in the ontological definition, modeling, and implementation of an addressing database that is flexible enough to accommodate the variety of concepts and address formats used worldwide, along with direct and indirect references to places. Second, in the definition of an indicator that is able to quantify the degree of certainty that could be reached when a user-given, semi-structured address is geocoded into a spatial position, as a function of the type and completeness of the available addressing data and of the geocoding method that has been employed. This indicator, which we call Geocoding Certainty Indicator (GCI), can be used as a threshold, beyond which the geocoded event should be left out of any statistical analysis, or as a weight that allows spatial analysis methods to reduce the influence of events that have been less reliably located. In order to support geocoding activities and the determination of the GCI, we propose a conceptual schema for addressing databases. The schema is flexible enough to accommodate a variety of addressing systems, at various levels of detail, and in different countries. Our intention is to depart from the usual geocoding strategy employed in commercial GIS products, which is usually limited to the average American or British address format. The schema also extends the notion of

C. A. Davis Jr (✉)
Instituto de Informática, Pontifícia Universidade Católica de Minas Gerais, Belo Horizonte, MG, Brazil
e-mail: clodoveu@pucminas.br

F. T. Fonseca
College of Information Sciences and Technology, The Pennsylvania State University,
University Park, PA, USA
e-mail: ffonseca@ist.psu.edu

postal address to something broader, including popular names for places, building names, reference places, and other concepts. This approach extends Simpson's and Yu's *Comput. Environ. Urban Syst.*, 27: 283–307, 2003 work on postal codes to records of any kind, including place names and loosely formatted addresses.

Keywords address geocoding · geographic information systems · spatial databases · certainty assessment · postal addresses

1 Introduction

Postal or urban addresses are the most common resource city dwellers use to convey geographic locations. Therefore, addresses are usually the most common reference to events and phenomena that take place in urban areas. As a result, most urban GIS applications, such as of transportation and transit, public health, public safety, and tax collection, are required to generate coordinates from addresses—as provided by the general public—in order to be able to visualize and analyze their data.

Addresses are also an important link to legacy systems, which contain data that are important for historical purposes, and for updating or populating spatial databases. Such legacy data, mostly alphanumeric in nature, are likely to refer to spatial locations using addresses, particularly in the case of urban applications. It is estimated that 80% of the information local governments use is associated to geographic locations, and most of those are related to addresses [5], [9].

The Web also presents a number of important and interesting challenges involving the recognition and location of addresses [2]. Addressing elements can be used as indirect references to the places referred to by the text contained in a Web page. Such references allow us to imagine search engines that are capable of not only answering the usual keyword-based queries, but also to locate pages connected to a specific place, thus enabling filtering pages down to a local context [14].

The importance of maintaining an up-to-date database of addresses is shown by traditional applications, such as emergency dispatching [7], where the time required to obtain a precise location from unstructured, sketchy, or even vague addressing data, provided by citizens in emergency situations, is critically important. However, the addressing database itself can be sketchy, incomplete, or imprecise. This is often the case in underdeveloped areas in emergent countries, for which data collection and updating can be a daunting task. In these cases, there is the need to produce the best possible approximation of a location based on the database contents.

In other application areas, time may not be as important as in emergency dispatching, but since data volumes tend to be high, efficiency in the generation of coordinates from addresses is also important. This is the case of records on criminal events and on public health issues [29]. Discussing the use of GIS in public health, Rushton [23] considers that it is also necessary to have tools that show the users the degree of *reliability* of the data (i.e., how sure can the user be that the location found actually corresponds to the indications provided as input). There is clearly a need for better links between public health records and addressing databases [21], as shown by the outbreak of severe acute respiratory syndrome (SARS). The SARS outbreak also motivates one of the topics of this paper, the issues involved in understanding the addressing systems in different countries. Further information on the different addressing systems in more than 193 countries can be found in The Global Sourcebook of Address Data Management [20].

Different geographic applications require different levels of accuracy regarding the location of the events or phenomena of their interest. This observation indicates that, while using the same addressing system with the same kinds of address-based location procedures, the quality of the results may suffice for some applications, but may be inadequate for others. Locating precisely each provided address in a short time is essential for emergency dispatching, while an epidemiologist may be content if each West Nile virus case is located only well enough for it to be associated with a reasonably small spatial unit, such as a Census tract. Thus, in this paper, we introduce a measurement to evaluate the degree of certainty associated with *geocoding*, i.e., the location of events or phenomena using addresses. This measurement, which we call *Geocoding Certainty Indicator* (GCI), shows how certain can the user be as to the actual position of the geocoded event. The resulting GCI can be used as a *threshold* for a filter, to leave out less reliable data, or as a *weight*, allowing the user to take the uncertainty into consideration during the spatial analysis process.[1]

In order to support geocoding activities and the determination of the GCI, we propose a conceptual schema for addressing databases. The schema is flexible enough to accommodate a variety of addressing systems, at various levels of detail, and in different countries. Our intention is to depart from the usual geocoding strategy employed in commercial GIS products, which is usually limited to the average American or British address format. The schema also extends the notion of postal address to something broader, including popular names for places, building names, reference places, and other concepts. This approach extends Simpson's and Yu's [25] work on postal codes to records of any kind, including place names and loosely formatted addresses.

The remainder of this paper is organized as follows. In Section 2 we present theoretical aspects of addresses, including a brief historical account and a set of concepts and formal definitions, leading to the design of a flexible addressing database. Section 3 presents a case study on international addresses, designed to review variations in different countries, and introduces the notion of an addressing location hierarchy. Section 4 presents the geocoding process, dividing it in three phases: approximation, matching and locating. Section 5 presents the determination of the GCI. Finally, Section 6 presents our conclusions and possibilities for future work.

2 A theory of addresses

Postal addresses are a common wayfinding resource, used in cities everywhere in the world. Even though addressing systems vary in a number of details [22], it is possible to look at the common points among addresses to conceive a data type for IS implementation. In conventional information systems, addresses are usually treated as attributes for some entities. In GIS, however, addresses are used as *locators* for urban events, and therefore should be modeled as separate entities, carrying along their own spatial location.

This view on addresses has produced a lot of work in *geocoding* (also called *address matching*) applications. These applications receive, as input, an address or a set of addresses, obtained from alphanumeric attributes corresponding to some fact which has been recorded in a database. These attributes are descriptors of locations, and are usually in

[1] This approach has been proposed by us in the past [6], but with a limited scope and with a preliminary, now outdated, formulation for the calculation of the indicator.

the form of postal addresses. The geocoding application tries to find a match for each of these descriptors in a reference database, which supposedly contains the locations for every relevant address, and, if successful, returns coordinates corresponding to its location [29]. In the case of roads and highways, in which addressing works as a linear referencing system [24], dynamic segmentation strategies have been proposed [11], associating a set of events to a linear feature through a distance along the line.

The combination of postal addresses and linear referencing should be enough to accommodate the needs of urban applications. However, in several parts of the world, the regularity one usually associates with addressing systems does not occur in practice. There are problems such as irregular or non-metric numbering, different criteria for naming streets, association of address numbering with regions instead of streets, and many more.

Postal addresses contain a number of components that can be seen as a spatial hierarchy that indicates, with increasing accuracy, the position of the mail recipient: a country name, a region/state name, a city name, a street name, and so on. Postal codes are usually employed as a shorthand for this hierarchy, a practical method to avoid human error in the interpretation of address components. We argue that, by taking this hierarchy into consideration, it is possible to obtain increasingly approximate positions for events associated with addresses, stopping at the point in which the needs of a given application are satisfied. With this approach, geocoding can return results with measurable accuracy and adequate to the needs of the application.

We have a broader view of addresses: urban addresses are more than just postal locators. Any place name that can be associated to a definite location, as recognized by the population in a local context, also constitutes a kind of urban address. In this sense, not only the formal components of addresses are acceptable as indications of a location. For someone who knows Washington, DC well enough, the string “The Mall, Washington DC” is fully sufficient to locate the corresponding place. Reference places, such as the names of monuments, stadiums, and parks, may also be recognizable and locatable by people, even though it is possible that no one resides or receives communications at those places. In a sense, we should be able to efficiently determine locations from any given urban address—in this broader view—if a minimum amount of context is available. Indirect references, such as telephone area codes, may also constitute valuable location indicators. In a way, postal codes are also indirect references, and in many cases the coding reflects a hierarchy of areas, something that may be useful if the intention is to quickly establish rough locations within a country [2]. This broader view of addresses is closer to the notion of a *gazetteer*, a sort of dictionary of place names [12], but geared towards urban places [26] and with varying degrees of accuracy.

2.1 A historical view on addresses

The idea of numbering buildings within cities arose from the need to guide visitors (or residents) about the location of a given dwelling or commercial activity [22]. The numbering system that most Westerners are accustomed to (i.e., sequentially increasing numbers along the street, with odd numbers to one side and even numbers on the other), is widely accepted, but is not dominant. Many cities implemented and maintain to this date different numbering systems for historical reasons, since standardization did not (or could not) occur everywhere.

The first addressing initiatives took place in Western Europe and China in the eighteenth century [22]. Numbering every building was not a general rule until government realized how a more efficient addressing would help cadastral and fiscal initiatives. Numbering in

Paris, one of the world's most advanced cities at that time, did not start until 1779, and was met with resistance from the population, especially from the dominant classes, who complained about being "equaled" to the lower social strata for being referred to by a simple number within a street.

One of the most important addressing systems in the western world is the metric numbering system, combined with the odd–even rule. Buildings are assigned numbers according to their metric distance from the beginning of the street, rounded up to the nearest odd or even number, or approximated in a way that every building gets a unique number. There are variations in which the numbering is sequential but block-oriented (for instance, assigning and distributing 100 numbers to a block), though the numbering is not distance-oriented. The metric system has the advantage of allowing an easy approximation of the distance between two addresses in the same street, while allowing for simple adaptation to new developments along each street. Similarly, some cities have a street naming (or numbering) which allows for a quick estimation as to the distance between two addresses.

There are other addressing and numbering systems in the world, which persist as a result of long-term usage and tradition, or by being adjusted to local needs and characteristics. The Japanese system, for instance, does not assign addresses according to consecutive numbers along a road, but numbers the houses according to their date of construction [15]. Most streets actually have no name at all, and, therefore, business cards typically show small maps printed on the back to indicate the location of a place. In Korea, numbers are assigned inside neighborhoods (called *dong*) within urban sectors (called *gu*), a hierarchy of areas that are named, not numbered [4]. In Kyoto, Japan, the Digital City project has been conceived with this kind of limitation in mind: a map-based user interface facilitates the location of points of interest for tourists and locals alike, since the addressing system seems to be too complicated to navigate without detailed mapping information [13].

Even though there is certainly a great variety of addressing systems throughout the world, there seem to be too few standardizing initiatives. In the United States, a standard for address data has been opened for public review in 2003 by the Federal Geographic Data Committee [10], as a proposal for the creation of a national spatial data infrastructure standard. It attempts to establish a basic terminology in order to create a semantic agreement regarding address data. For instance, the proposal defines addresses as "locators to places where a person or organization may reside or receive communications, but excluding electronic communications." This initiative is still under review. The UK has introduced British Standard 7666 to help in the development of a national framework for geocoding land and property information [18].

We observe that, even though addressing systems and local customs vary, some basic notions are present in most cases. Based on this observation, we propose in the next section a set of concepts for address components.

2.2 Addressing concepts

Addresses work as descriptions used by people to communicate positions and locations. These descriptions are composed based on knowledge that is common to the originator and the receiver of the communication, usually assuming a specific context. In postal addresses, the sender cannot assume the existence of such a context or of common knowledge between himself and the postal workers that will route and deliver the package; therefore, postal addresses are the most structured form of addressing people use. In other situations, recognizing locations from descriptions depends on some knowledge on the part of the receiver.

Addresses can be either *direct* or *indirect* references to places. Direct references provide a structured description, such as a postal address, or a definite place name, while indirect references comprise numbers or codes that refer to a location through some previously created relation. Examples of indirect references include telephone area codes, highway exit numbers, some types of postal codes, and cadastral codes.

Direct addresses can be *absolute*, i.e., references to a definite place, or *relative*, i.e., indications to some place positioned in the vicinity of a reference location. Relative addresses usually take the form of an absolute reference attached to an indication of relative positioning, such as "100 km to the North of Paris" or "close to the Ambassador Hotel". The indication of relative positioning is usually formed by an expression that denotes a spatial relationship (e.g., near, close to, beside, *x* km/miles from, *x* minutes from, and so on) and an absolute address [2].

We will not discuss relative or indirect addresses further, since our focus here is on locating direct absolute addresses. We recognize the following types of direct absolute addresses:

- *Postal addresses:* Structured descriptions, containing a hierarchy of places (e.g., country, state, city, neighborhood, street) and complementary information (building number, building name, apartment number) used to pinpoint a specific location. Postal codes work as both a shorthand for the hierarchy of places and a redundant item, against which other elements can be checked.
- *Linear references:* Used in the identification of places along roads or railways, are formed by a distance along a linear feature, considering a conventional start point ("mile marker 129, U.S. Route 66") or by a distance from a given point ("100 km from the Mexican border, along Highway 123") [24].
- *Place names or toponyms:* Names of known places, either natural or man-made (such as buildings), usually context-dependent, used by people by themselves or as landmarks. Examples: "the Eiffel Tower," "Manhattan."
- *Composite addresses:* Place indications composed by a place name and some complementary information, such as distance and/or direction, from which to indicate the location of something nearby, or by the combination of two or more place names. Example: "at the corner of Oak St. and First Avenue."

Depending on the type of address used, an accurate position can or cannot be obtained. Many of the addressing types express a rather general position, and are used with the intention of providing a rough idea about some location. We show later how to quantify this uncertainty, so that it can be taken into consideration in analysis, rankings, or filters in the geocoding of related events.

In this paper, we are particularly interested in urban addresses, and therefore we will look deeper into postal addresses and urban place names, leaving the other categories for future work on context-dependent addressing. For now, we consider the following set of concepts on addressing:

- *Thoroughfare.* A named public space, usually associated with the concept of a passage within a city; a generic concept including the notions of *street, avenue, plaza, square, road, alley, lane*, and *boulevard* (*thoroughfare types*).
- *Thoroughfare name.* The name officially or popularly associated with a thoroughfare, which is usually characterized also by the thoroughfare type.
- *Crossing.* The location at which two or more thoroughfares meet (crossroads, street crossing).

- *Building number.* The number used to identify a building at a thoroughfare. Does not include the identification of property units within a building, such as apartment, room, office, suite, unit, floor or others (*complement*).
- *City sector.* A named division of a municipality's territory, recognizable by people as a definite place or region (alternative names: *subdivision, district, borough*, or *gu* in Korea); may include several neighborhoods.
- *Neighborhood.* A subdivision of a municipality's territory, within which dwellers might consider themselves to be neighbors. Neighborhoods are named, often after the real estate development project from which they were created (alternative names: *region*, or *dong* in Korea)
- *City.* A collection of human dwellings, an urban area. Depending on the size and local custom, it may assume different denominations (*village, town, metropolis, municipality*).
- *State.* A region of a country (mainly in the federal system), usually ruled by a regional government, comprising several municipal areas.
- *Postal Code.* An alphanumeric code used by postal authorities to facilitate the sorting of mail in preparation for its distribution by a mail carrier.
- *Landmark.* A place whose name is well known by the population, usually serving for routing or orientation. Landmarks include human constructions (e.g., buildings, stadiums, monuments, and bridges) as well as distinct natural landmarks (e.g., *Niagara Falls, Matterhorn*).

With these basic notions in mind, we define the concept of an address as follows:

- *Address.* A description, including the names and any complementary pieces of information, which allows someone to uniquely identify a place.

Therefore, addresses are formed by combining these basic elements, provided there are enough elements to determine a unique location. The exact elements that are used to compose an address, along with the sequence and arrangement of these elements in the description, vary around the world.

We will now proceed to the definition of a flexible database schema for an addressing database, in an effort to accommodate all kinds of addressing systems. This schema is the basis of a comprehensive geocoding application, which will be described later.

2.3 Conceptual modeling

A spatial database, designed to allow someone to obtain a location from an arbitrary address description, must store all the elements of an addressing system along with their locations. Populating and maintaining such a database up-to-date can be a costly and complex task.

Considering this difficulty and reflecting on the needs of applications that need to locate large volumes of addresses in underdeveloped areas, our goal is to create a schema that is as flexible as possible, avoiding rigid domain constraints and mandatory attributes. Thus, the database can be useful even while it is partially populated. Of course, more detailed and complete data will result in more reliably geocoded locations (as indicated by the GCI), but the user should be able to refine the database contents incrementally. Souza et al. [26] shows a study in which, from analyzing a set of Web documents, a listing of the most frequently mentioned locations is obtained. Eventually, these locations should be added to the addressing database.

Figure 1 presents the conceptual schema for the addressing database, developed using OMT-G [3], an extension of the Universal Modeling Language (UML) dedicated to geographic applications. In the OMT-G notation of each georeferenced class there is a pictogram which indicates the spatial representation type that is to be employed. Simple

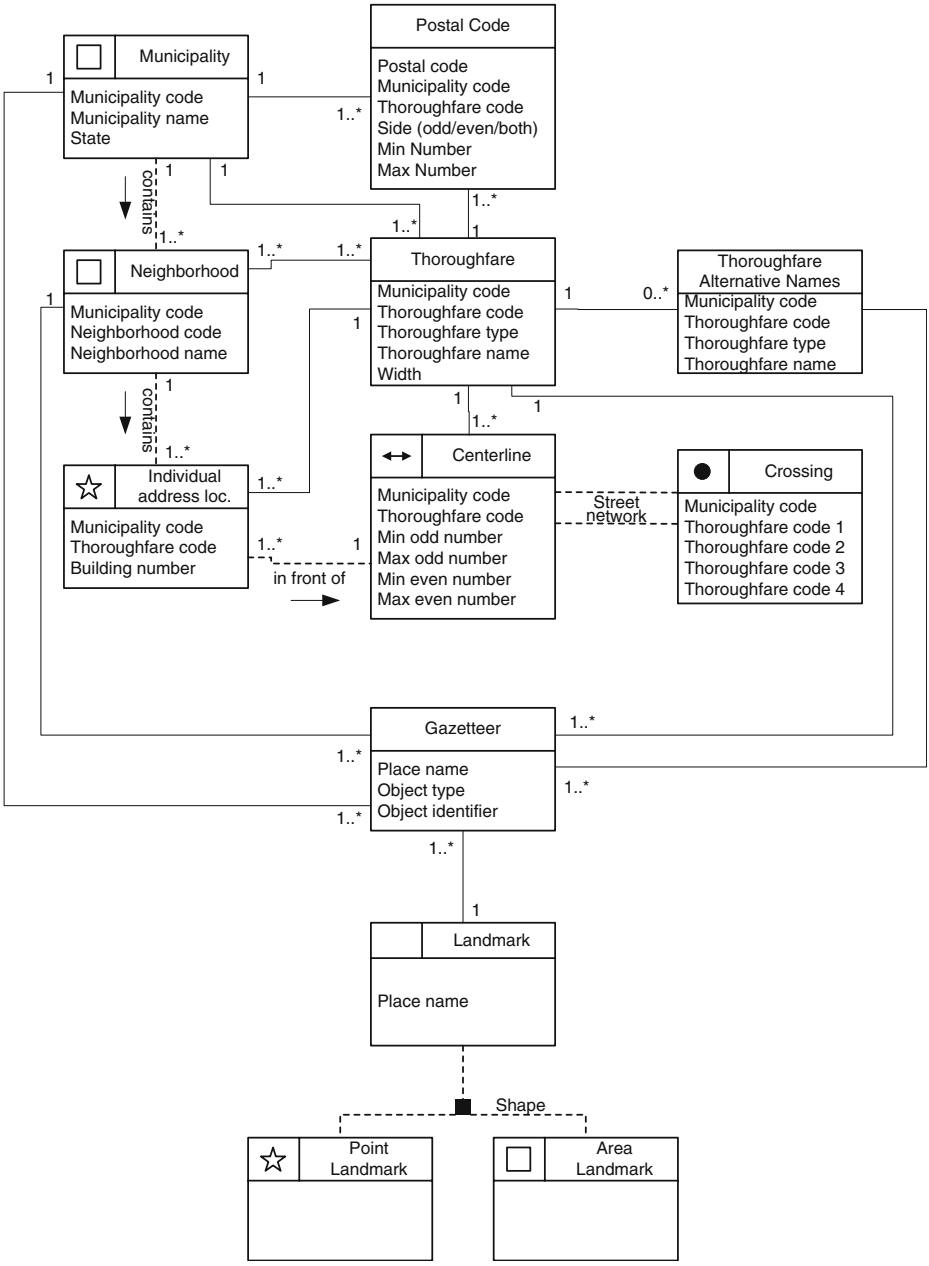

Fig. 1 OMT-G Addressing schema—Class diagram

associations are denoted with continuous lines, and spatial relations are denoted with dashed lines.

In the schema, we included a class for individual addresses, represented as points. With this approach, we are able to avoid situations in which irregular numbering (e.g., numbers that are not always increasing along the thoroughfare, or that disregard the usual odd–even side rule) or unusual addressing systems (such as the Japanese and the Korean ones) prevent the use of numbering ranges in each block. In the schema, such addresses are defined in the `Individual address location` class, which contain a thoroughfare code and a building number. We also use points to represent landmarks, i.e., places that are identified by name, and that are precisely locatable, such as monuments or buildings (`Point Landmark`). In case the space corresponding to the landmark is significantly large, such as in a park or a sports arena, a polygonal representation is used (`Area Landmark`).

The `Neighborhood` class allows for the representation of any intra-urban spatial reference units, including districts and boroughs, as polygonal objects. The `Centerline` class allows for the geometric materialization of thoroughfare segments, and may contain information on the ranges of building numbers at each side (odd and even ranges), in the manner of the Topologically Integrated Geographic Encoding and Referencing (TIGER) files [27]. General information on the thoroughfares themselves is included in the `Thoroughfare` class. The `Thoroughfare alias` class stores historical names and popular nicknames to certain thoroughfares. Each `Centerline` object is related to two Crossing objects, thus forming a street network. Using centerlines and crossings allows the location of places described as, for instance, “at the corner of Oak and Main St.” Postal codes are also included, considering both the association of several codes to the same street (defined in ranges) and the association of a single code to the entire municipality. Postal codes play the role of disambiguators in case several thoroughfares exist with the same name. The schema yields the definition of a *gazetteer* [15], which will receive place names from all other classes and, therefore, can be used to determine the possible nature of a place, given an unqualified name.

Direction indicators associated with street names—a common setting in U.S. cities—can become part of the street name. Therefore, streets that are divided by a central avenue, having independent numbering in each direction, should be modeled as two separate streets. For instance, the Washington, DC branch of Greenpeace is located at `702 NW H Street`; in the addressing database, the building number is `702`, the thoroughfare type is `Street` and the thoroughfare name is `NW H`. The other half of this street, namely `NE H`, is encoded as a separate thoroughfare.

From the conceptual schema in Fig. 1, we generated an object-relational schema (Fig. 2), which is used in the following sections to present the necessary geocoding actions in a more detailed way. In the object-relational schema, we denoted as `GEOM_Point`, `GEOM_Line`, or `GEOM_Poly` the geometric representation associated to each georeferenced object class found in the conceptual schema. Using the contents of this database, a geocoding system has been built, as presented in Section 4.

3 Common addressing concepts: A brief case study

In order to verify whether the set of concepts presented earlier is broad enough, and to reinforce our argument in favor of the kinds of resources defined in this paper, consider the case of analyzing and treating addresses from the list of offices maintained by a worldwide

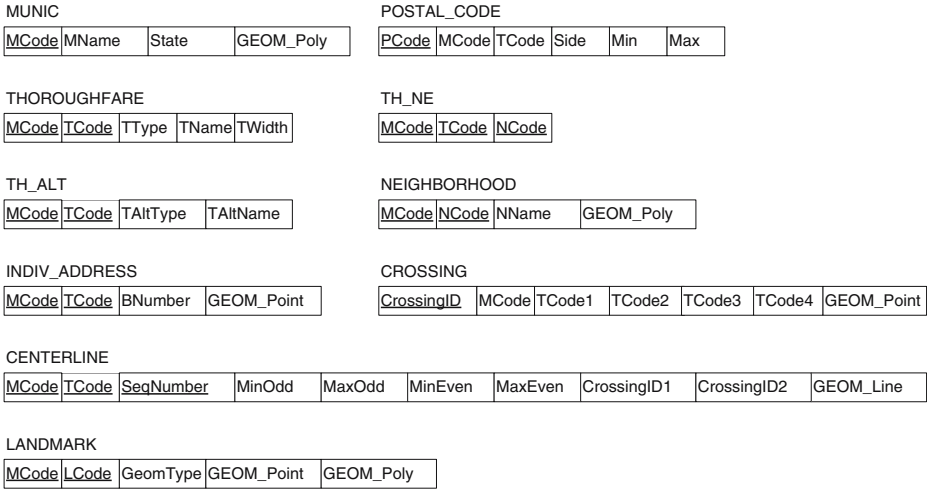

Fig. 2 Object-relational schema

organization, such as Greenpeace. Table 1 lists 20 addresses of Greenpeace offices,[2] each one in a different country. We obviously did not choose postal addresses based on a P.O. box or the like, since we are interested in the actual structure of each address and in the place names contained in them. Table 2 shows the same addresses, separating their basic components, according to our model (thoroughfare type, thoroughfare name, building number, building name, neighborhood, city, state, country, postal code, and complement). In this compilation, no address includes all of these components and no component is used by all addresses. Names of famous places are employed as addresses: for instance, the Mexican branch of Greenpeace is located at a street or avenue named Andalucia, which is also the name of a Spanish region. Such ambiguities make current Web search engines inefficient when we attempt to locate pages that refer to specific geographic locations [2]. What we called *states* in the example and in the remainder of the paper may refer to any hierarchical level between *city* and *country*, such as counties, provinces or territories.

Assuming that there is no universal standard, we can now point out some common traits among all these addressing systems, considering the intention to assign coordinates to every address that can be recognized by postal authorities. The concepts of *street* (or, more generally, *thoroughfare*), *building number* (or *name/identifier*), *neighborhood*, and *city* or *municipality* seem to be approximately the same all around, even though in some situations concepts of a more cadastral nature, such as *block*, are used as address references. *Postal codes* are also useful in addresses, since they are widely used by the population, even though they take on different formats in each part of the world. Incomplete, inaccurate, or hard-to-use addresses can be associated to *indirect references*, which can be thought of as distinct landmarks within the city or points that are widely known and recognized by the public. These references can be thought of as points, in case their dimensions are small, or as areas, in case their name is associated with a wide piece of land.

[2] Extracted from http://www.greenpeace.org/international_en/contact/index-int

Table 1 Greenpeace postal addresses around the world

Number	Postal addresses
1	Mansilla 3064, 1425 Buenos Aires, Argentina
2	Siebenbrunnengasse 44, A-1050 Vienna, Austria
3	Haachtsesteenweg 159 — 1030 Brussels, Belgium
4	Rua Alvarenga, 2331, Butanta 05509-006, Sao Paulo/SP, Brazil
5	250 Dundas Street West, Suite 605, Toronto, Ontario M5T 2Z5 Canada
6	Eleodoro Flores 2424, Ñuñoa, Santiago, Chile
7	1/F, Tung Lee Commercial Building, 95 Jervois Street, Sheung Wan, Hong Kong, China
8	First Floor, Old Town Hall, Victoria Parade, Suva, Fiji Islandsisra
9	22 rue des rasselins 75020 Paris, France
10	Grosse Elbstrasse 39, D-22767 Hamburg, Germany
11	Zoodochou Pigis 52c, GR-106 81 Athens, Greece
12	3360, 13th B Main, HAL II Stage, Indiranangar, Bangalore — 560 038, India
13	Viale Manlio Gelsomini 28, 00153 Rome, Italy
14	N F BLDG. 2F 8-13-11 Nishishinjuku, Shinjuku-ku, TOKYO 160-0023 Japan
15	Andalucia 218 Col. Alamos, Mexico D.F. CP 03400, Mexico
16	113 Valley Road, Mount Eden, Auckland, New Zealand
17	Unit 326 Eagle Court Condominium 26, Matalino Street, Diliman, Quezon City, Philippines
18	San Bernardo 107, 28015 Madrid, Spain
19	Canonbury Villas, London N1 2PN, United Kingdom
20	702 H Street NW, Suite 300, Washington DC 20001, USA

As a conclusion, we observe that we can treat addresses as abstract data types, in which a subset of the components are required to determine the location univocally. The role of each component varies in the address. There are:

- Components that indicate the location only when used as an integrated set; for instance, building numbers are meaningless by themselves, they must always be associated with a thoroughfare;
- Components which establish a refinement over the location provided by other components or a set of components, as in the case of complements;
- Components that hierarchically approximate the location, such as postal codes;
- Components that indicate the location given some context, such as building names or neighborhoods, for which it is often necessary to establish the city in which they exist, for disambiguation.

Considering the semantics of the addressing system and its components as presented earlier, and observing how addressing works in several different parts of the world, we can infer a hierarchy of successive approximations to the geographic location of the address (Fig. 3).

Using a database modeled as proposed in this paper, and considering the hierarchy in Fig. 3, applications can be developed so that (1) geocoding can stop if the required degree of accuracy is met, or (2) a location can be provided even in the absence of some addressing components. As an example of the first case, consider a geomarketing application with nationwide coverage, for which it is sufficient to identify the city of residence of each registered customer, in order to determine the ideal location for a new store. In the second case, consider a public health system on epidemiology, which needs to locate as precisely as possible the residence of people that have been infected with a contagious disease. In consolidated areas of the city, the location is obtained through individual addresses; in

Table 2 Parsed Greenpeace postal addresses

Th. type	Thoroughfare name	Building number	Building name	Neighborhood	City	State	Country	Postal code	Complement
	Mansilla	3064			Buenos Aires		Argentina	1425	
Gasse	Siebenbrunnen	44			Viena		Austria	A-1050	
	Haachtsesteenweg	159			Brussels		Belgium	1030	
Rua	Alvarenga	2331		Butantã	São Paulo	SP	Brazil	05509-006	
Street	Dundas West	250			Toronto	Ontario	Canada	M5T 2Z5	Suite 605
	Eleodoro Flores	2424		Ñuñoa	Santiago		Chile		
Street	Jervois	95	Tung Lee commercial building	Sheung Wan	Hong Kong		China		1/F
Parade	Victoria		Old town hall		Suva		Fiji Islandisra		First Floor
Rue	Des Rasselins	22			Paris		France		
Strasse	Grosse Elb	39			Hamburg		Germany	D-22767	
	Zoodochou Pigis	52c			Athens		Greece	GR-106 81	
	13th B Main	3360	HAL II stage	Indiranangar	Bangalore		India	560038	
Viale	Manlio Gelsomini	28			Rome		Italy	00153	
	Nishishinjuku (region)	8-13-11	N F bldg	Shinjuku-ku (subregion)	Tokyo		Japan	160-0023	
	Andalucia	218		Col. Alamos	Mexico	Mexico D.F.	Mexico	03400	
Road	Valley	113		Mount Eden	Auckland		New Zealand		
Street	Matalino		Diliman	Eagle Court Condominium 26	Quezon City		Philippines		Unit 326
	San Bernardo	107			Madrid		Spain	28015	
			Canonbury villas		London		United Kingdom	N1 2PN	
Street	NW H	702			Washington	DC	USA	20001	Suite 300

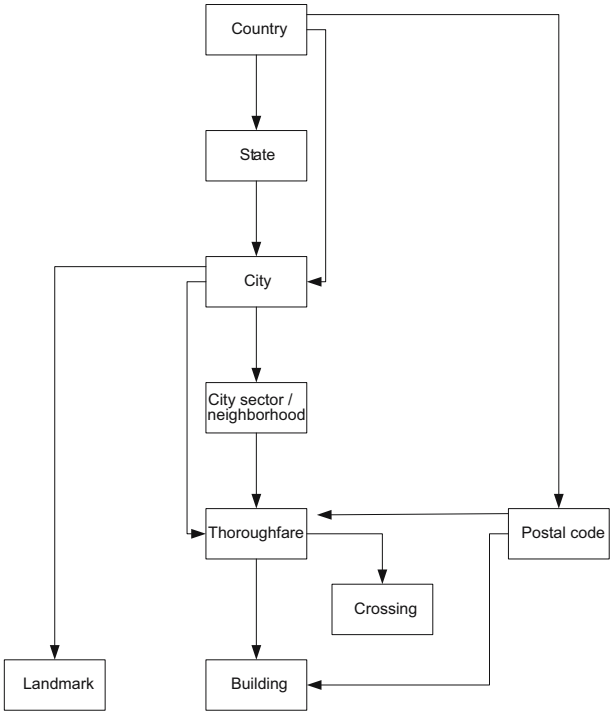

Fig. 3 Addressing concepts hierarchy

recent developments, for which addressing information is not detailed enough, the location is approximated using neighborhood limits.

In order to be able to recognize and locate addresses such as the ones presented in this section, databases containing the place names and their location are required. Addressing databases are usually available in such countries as part of a national information infrastructure strategy, presenting high quality and low cost. The foremost example of this kind of information are Topologically Integrated Geographic Encoding and Referencing (TIGER) files [27]. Private sector companies have access to this material, and invest in its improvement, thus demonstrating that address databases can be a valuable economic asset. In the UK, the Ordnance Survey produces and sells licenses on an address point database that contains over 25 million locations, along with a coordinate list for the 1.6 million distinct postal codes within the country [16]. Even if an addressing database is available, there are many cities in which the addressing system is considerably different from the American and British cases, and thus the geocoding methods provided by commercial GIS packages will not work as expected.

Emergent countries, such as Brazil and India, usually do not have such a complete and organized addressing database from which to accurately and quickly generate positions from addresses. The consequences of this for urban geographic applications are manifold, since georeferencing point data can take much longer, resulting in poor data quality from consistency and precision problems. Furthermore, large cities in emergent countries often contain slums, shantytowns, and other types of low-income areas that are characterized by irregular occupation, and often in these areas there is not even an address plaque at each dwelling. Also, in many cases the addressing database is not as complete as it

should be, due to lack of information or to the cost of generating and maintaining a detailed database in places where fast and chaotic growth, and irregular land occupation, are predominant.

The usefulness of georeferenced addressing databases is such that, in many places, local government departments and infrastructure service providers constantly invest in their creation and maintenance, using information from alphanumeric cadastres and conventional cartographic sources. Since there is often no established standard for the creation of such information resources, regional or national efforts that need to work with massive amounts of point-georeferenced data, especially in fields such as epidemiology and crime fighting, are severely hindered [17]. We now proceed to the definition of geocoding tools and techniques based on the proposed database schema, and considering the hierarchy of addressing concepts.

4 Geocoding

The determination of a geographic position from a descriptive address is called *geocoding*. Geocoding tools are important components of urban GIS. These tools usually comprise two interdependent parts: a set of *geocoding methods* and the *addressing database*, sometimes called *reference database*.

Geocoding methods are a set of algorithmic procedures developed for the purpose of comparing a supplied list of textual addresses, each associated with some event or phenomenon of interest, to the contents of an addressing database, considering a standardized data model. Many commercial GIS packages include geocoding tools, but their algorithms require the addressing database to be structured according to a specific schema, usually following the characteristics of the addressing system used in some developed countries, particularly the United States and the United Kingdom.

Generically, geocoding works in three stages (Fig. 4). First, the input address data must be analyzed and structured according to some template, in a *parsing* stage. This first stage may be unnecessary, if the input addresses are already structured in a convenient way; this is often the case when addresses come from a legacy information system. The second stage (*matching*) compares the structured address data with the contents of the addressing database, using several different geocoding methods. In the third geocoding stage (*locating*), a location is determined according to the results of the matching stage. In our methods, a certainty indicator is calculated for each stage, and a general indicator is calculated at the end. Each of these stages is described in the next sections.

4.1 Parsing

This stage receives an unstructured string of text containing an address. The objective of this stage is to create a tuple containing every significant piece of information from the original address string. If necessary, addressing elements found in the string must be normalized or adjusted before becoming fields in the tuple. This process is called the *parsing* of the original address. In GIS, the process of recognizing geographic context from geographic data sets is specifically referred to as *geoparsing* [16].

The parsing stage is further complicated by the fact that each different country (or sometimes city) may have a different standard for the presentation of addresses, as

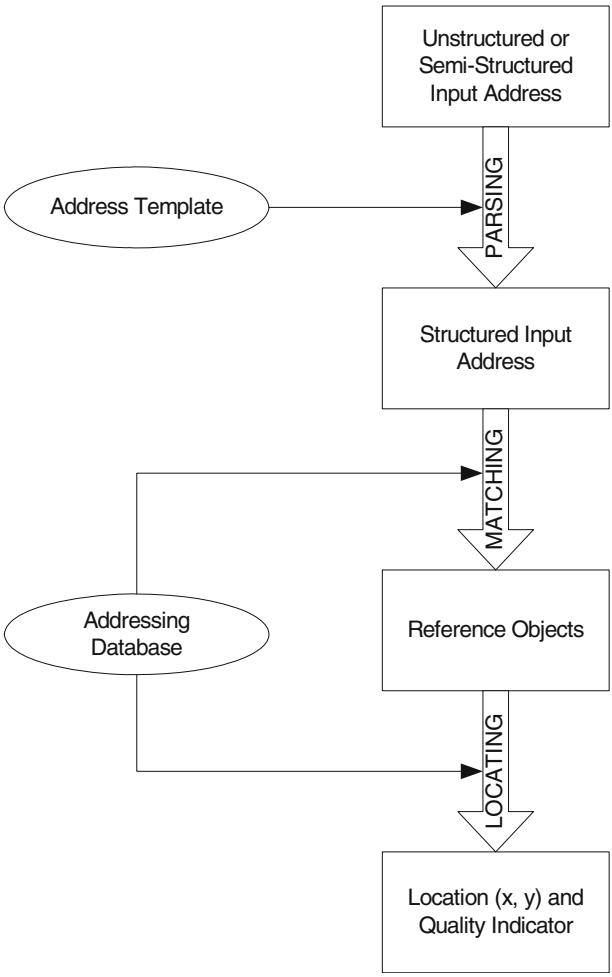

Fig. 4 General geocoding schema

demonstrated in the Greenpeace example. We must, therefore, be able to supply an *address template* for each situation, which indicates how the parser should interpret each part of the supplied addresses. A template for American postal addresses would look like this:

```
<building number> [<direction indicator>]
<thoroughfare name> <thoroughfare type> [<complement>]
<city> <state> <postal code>
```

A typical Brazilian postal address, on the other hand, would be defined like this:

```
<thoroughfare type> <thoroughfare name>
<building number> [<complement>] [<neighborhood>]
<postal code> <city> <state>
```

The definition of such a template also helps solving situations in which the input addresses already present some structure. It is quite common, for instance, that fields such as city,

state, and postal code are encoded separately from the main address field in legacy information systems, leaving us with only the thoroughfare identification and numbering to deal with. Depending on the source, therefore, we may need different parsing techniques.

The algorithms that can be used for parsing the addresses are similar to those used in programming languages in order to assess the syntax of a language construct. The string gets initially divided into tokens, considering a set of white space characters (blanks, commas, points, hyphens, and so on) as delimiters. The resulting tokens are then analyzed sequentially, in an attempt to determine the function of each one of them. The analysis of each token can use the addressing database, if it is necessary to establish hypotheses as to what is the correct interpretation of each term (token) in the address. In this process, the gazetteer defined in the addressing database schema can be helpful. Through the use of the gazetteer, and by scanning the template in order, it is possible to decide the adequate interpretation on situations in which the same name is related, for instance, to a municipality and to a state (for instance, "São Paulo" in Brazil is both a city and a state, and possibly the name of many streets all over the country). This approach contrasts with the parsing of free-form text for information retrieval (IR) [1], since in that case the intention is mostly to obtain context information by analyzing the presence, sequencing and frequency of keywords. The techniques we implemented contain a mix of language parsing with information retrieval parsing.

The result of the parsing process is a tuple, containing a fully structured address, in which there is an attribute for each of the addressing components that are required in the matching stage. Specifically, we are looking for attributes such as those presented in the Greenpeace example, and thus the tuples have the following structure:

```
<Ttype, Tname, Bnumber, Bname, Nname, City,
        State, Pcode, Complement>
```

where Ttype is the thoroughfare type, Tname is the thoroughfare name, Bnumber is the building number, Nname is the neighborhood name, Bname is the building or landmark name, and Pcode is the postal code.

As in the case of the Greenpeace example, not all attributes are present every time; therefore, the matching stage must be able to deal with null values, thus limiting the attributes that can be used to find matching geographic objects.

4.2 Matching

After obtaining a structured address, the matching stage consists in determining the objects in the addressing database that correspond to the supplied data. A match is attempted for each addressing component, with the objective of determining a single object from the addressing database for each significant part of the address (city/state, postal code, neighborhood, thoroughfare, building name or landmark, individual address). If multiple matches are found, a disambiguation procedure is used, possibly employing previously matched components. If the disambiguation is not possible, or if no match is found, no object is returned. From the (complete or incomplete) set of objects that result from the matching, the locating phase will determine the location from the most adequate one, i.e., the one for which there is less uncertainty.

Matching occurs by comparing place names and other data contained in the address tuple with attributes of the object classes that compose the addressing database. There are three major problems to be dealt with in this phase. First, there is the issue of abbreviations, something that can occur in thoroughfare types (*Ave*. for *Avenue*, for instance) and in place

names of every sort (*St. Patrick* instead of *Saint Patrick, Franklin D. Roosevelt* or *FDR* instead of *Franklin Delano Roosevelt*). Second, there are always spelling mistakes and other kinds of imprecisions in place names. Finally, there is the need to perform disambiguation in the case of multiple matches.

To solve the abbreviations problem, a pre-processing phase may replace occurrences in attributes such as thoroughfare type using a list of common abbreviations. Unusual abbreviations or abbreviations in thoroughfare names may be treated as alternative names, and included in the `Thoroughfare Alias` class of the addressing database. Spelling problems can be countered using approximate string matching algorithms, such as *Levenshtein distance* [19] and *Shift-And* [28]. Soundex-like procedures can also be used, but we decided against including them in our implementation, based on the results presented in [30], and on our perception that phonetic methods would be best employed if source data were primarily in English, without many personal names or foreign words, which is clearly not the case of addressing data. Approximate string matching algorithms are able to find matches even when there are differences, such as missing characters, inverted characters, or extraneous characters, between the given string and the matching pattern. These algorithms are adapted to multilingual support, by granting the equivalence between accented characters and their canonical counterparts (for instance, "á" and "ä" are considered equivalent to "a" for approximate matching purposes). If multiple matches result from this process, additional information must be used, when available. For instance, if there are two streets with the same name in given city (a common situation in developing countries), the neighborhood name or the postal code can be used to select a single result. If this is impossible, for any reason, we consider that the matching has *failed*. The matching also fails if no match is found with the available data.

Names and numbers can be matched exactly or approximately. Names are matched using exact or approximate string matching algorithms, as previously mentioned. In the case of approximate matching, an upper bound on the number of differences that can occur between two strings that are considered to match has to be established. In most situations, we used a limit of 20% of the number of characters in the pattern (i.e., the string for which a match is being sought). Notice that we always test for exact matches first, and then look for approximate matches. For that reason, we do not employ the usual information retrieval technique, by which strings are converted to a canonical form prior to their comparison to a reference, which is also kept in canonical form.

Considering the frequency of personal names which are assigned to thoroughfares, regions, and cities, we implemented special variations of approximate string matching routines by which the abbreviation of intermediate names can be considered a full match. Our routines also consider the possible inversion in the sequence of names, establishing a lighter penalty for such cases as compared to non-matching names.

In the case of numeric matching, the approximation can either be numeric (mathematical difference between two numbers) or string-based (number of differences between strings corresponding to each number). From this point forward, exact string or numeric matching will be denoted using the "=" operator, while approximate matching will be denoted using a generic `Match` function. `Match` takes, as arguments, two strings and a real number, indicating the upper bound on differences as a percentage of the length of the first string, and returns a Boolean value, indicating whether the strings are considered to match or not.

Considering the addressing schema and the tuple resulting from the parsing stage, we propose the matching operations listed in Table 3. There is a group of possible matching operations for each possible resulting object, namely objects that belong to the `Postal code`, `Municipality`, `Neighborhood`, `Thoroughfare`, `Centerline`, `Individual`

Table 3 Matching operations

Op number	Operation	Resulting object
1.1	Find a POSTAL_CODE object such that t.Pcode = POSTAL_CODE.Pcode	OPC
2.1	Find a MUNIC object such that Match(t.City, MUNIC.Mname, 20%) = TRUE	OMU
2.2	Find a MUNIC object such that Match(t.City, MUNIC.Mname, 20%) = TRUE and t.State = MUNIC.State	OMU
2.3	If OPC is not null, find a MUNIC object such that MUNIC.Mcode = OPC.Mcode	OMU
3.1	Find a NEIGHBORHOOD object such that Match(t.Nname, NEIGHBORHOOD.Nname, 20%)	ONE
3.2	Find a NEIGHBORHOOD object such that Match(t.Nname, NEIGHBORHOOD.Nname, 20%) = TRUE and OMU.Mcode = NEIGHBORHOOD.Mcode	ONE
4.1	Find a LANDMARK object such that Match(t.Bname, LANDMARK.Lname, 20%) = TRUE	OLM
4.2	Find a LANDMARK object such that Match(t.Complement, LANDMARK.Lname, 20%) = TRUE	OLM
5.1	Find a THOROUGHFARE object such that Match(t.Tname, THOROUGHFARE.Tname, 20%) = TRUE	OTH
5.2	Find a THOROUGHFARE object such that Match(t.Tname, THOROUGHFARE.Tname, 20%) = TRUE and t.Ttype = THOROUGHFARE.Ttype	OTH
5.3	Find a THOROUGHFARE object such that Match(t.Tname, THOROUGHFARE.Tname, 20%) = TRUE and OMU.Mcode = THOROUGHFARE.Mcode	OTH
5.4	Find a THOROUGHFARE object such that Match(t.Tname, THOROUGHFARE.Tname, 20%) = TRUE and t.Ttype = THOROUGHFARE.Ttype and OMU.Mcode = THOROUGHFARE.Mcode	OTH
5.5	Find a TH_ALT object such that Match(t.Tname, TH_ALT.Taltname, 20%) = TRUE	OTH
5.6	Find a TH_ALT object such that Match(t.Tname, TH_ALT.Taltname, 20%) = TRUE and t.Ttype = TH_ALT.Talttype	OTH
5.7	Find a TH_ALT object such that Match(t.Tname, TH_ALT.Tname, 20%) = TRUE and OMU.Mcode = TH_ALT.Mcode	OTH
5.8	Find a TH_ALT object such that Match(t.Tname, TH_ALT.Taltname, 20%) = TRUE and t.Ttype = TH_ALT.Talttype and OMU.Mcode = TH_ALT.Mcode	OTH
5.9	If OPC is not null, find a THOROUGHFARE object such that OPC.Tcode = THOROUGHFARE.Tcode	OTH
6.1	If OTH is not null, find an INDIV_ADDRESS object such that OTH.Tcode = INDIV_ADDRESS.Tcode and t.Bnumber = INDIV_ADDRESS.Bnumber	OIA
6.2	If OTH is not null, find an INDIV_ADDRESS object such that OTH.Tcode = INDIV_ADDRESS.Tcode and | t.Bnumber−INDIV_ADDRESS.Bnumber| < tolerance	OIA
6.3	If OTH is not null, find an INDIV_ADDRESS object such that OTH.Tcode = INDIV_ADDRESS.Tcode and Match(str(t.Bnumber), str(INDIV_ADDRESS.Bnumber), 20%) = TRUE	OIA
6.4	If OTH is not null, find an INDIV_ADDRESS object such that OTH.Tcode = INDIV_ADDRESS.Tcode and | t.Bnumber−INDIV_ADDRESS.Bnumber| is minimal	OIA
7.1	If OTH is not null, find a CENTERLINE object such that OTH.Tcode = CENTERLINE.Tcode and (if t.Bnumber is odd t.Bnumber <= CENTERLINE.maxodd and t.Bnumber >= CENTERLINE.minodd) or (if t.Bnumber is even t.Bnumber <= CENTERLINE.maxeven and t.Bnumber >= CENTERLINE.mineven)	OCE
7.2	If OTH is not null, and t.Bnumber does not fit into any numbering ranges from corresponding CENTERLINE objects, return a temporary CENTERLINE object such that OCE.Geom_line = GEOMETRIC_UNION (SELECT all CENTERLINE where OTH.Tcode = CENTERLINE.Tcode) and OCE.Tcode = OTH.Tcode	OCE

address, and Landmark classes. Table 3 indicates such object instances respectively as OPC, OMU, ONE, OTH, OCE, OIA, and OLM to keep the notation short.

The operations in each group are executed in sequence, stopping at the first successful match. For instance, if a match is attempted on a thoroughfare name, operations 5.1 to 5.9 are successively executed. If a match is found after, say, operation 5.3, then the OTH object receives a copy of the matching object from the addressing database, and all remaining operations in group 5 are bypassed. If there is no match after the nine operations, then the OTH object receives a *null* value.

Even though the operations in each group are performed in sequence, sometimes some of the operations can be dismissed. For instance, if there is no information on alternative names for thoroughfares in the database, the operations involving such data do not need to be executed. Likewise, if the input data do not include postal codes, matching actions involving such data can be disregarded.

4.3 Locating

After the set of objects resulting from the matching phase is obtained, the locating phase must analyze their contents to determine the most accurate coordinate for the given address. There are seven objects (OPC, OMU, ONE, OLM, OTH, OIA, OCE), any of which can either contain a copy of a valid object, or a null value. However, only OMU, ONE, OLM, OIA and OCE are associated to some geometry, from which a location can be determined.

There are three different methods to determine a coordinate from these objects. The most accurate one is to copy the coordinates from a point object. The second one consists in interpolating a coordinate along a line object, considering the (min, max) numbering range along a centerline object. The third one requires the determination of an arbitrary coordinate within a line or area object. This can be done either by choosing the object's centroid, or by determining a random point along the line or inside the object's area. We prefer the latter method, since it avoids the overlapping of possibly many resulting coordinates in a single point, and because we think that, if we cannot determine the location more precisely, any point within the object should be equally probable.

In the case of point-based locating, we can either use the coordinates of OIA or the coordinates of OLM. The first is preferred, since this indicates that a complete address has been found. Line-based locating occurs only if no point-based locating is possible, and if OCE is valid. Invalid numbering ranges (as in method 7.2 in Table 3) require a random point generation along the OCE object. Area-based locating occurs only if none of the previous methods has been possible. Coordinates generated from ONE are preferred, since neighborhoods are obviously smaller than the municipalities. We resort to random points within the municipality only if there is no other option.

Of course, of the results from the matching phase do not allow the determination of any coordinate, the locating phase *fails*. Failure in this phase indicates that the location indications contained in the supplied address were either insufficient or unrecognizable, considering the contents of the addressing database.

The algorithm below summarizes the locating phase in pseudocode:

```
IF OIA is valid THEN
  Location = PointCoord (OIA)

ELSE IF OLM is valid THEN
  Location = PointCoord (OLM)
```

```
ELSE IF OCE is valid THEN
IF OCE range is valid then
  Location = RangeInterpolation (OCE, t.Bnumber)

ELSE
  Location = RandomPoint (OCE)

ELSE IF ONE is valid then
  Location = RandomPoint (ONE)

ELSE IF OMU is valid THEN
  Location = RandomPoint (OMU);

RETURN (Location);
```

5 A geocoding certainty indicator

Different applications in GIS make different uses of addresses. The need of more or less accuracy in the use of georeferenced information may depend on scale of presentation, the accuracy of source data, and of the kind of spatial analyses that are intended. Any practitioner in this field certainly recognizes that data (both supplied by the application and included in the reference database) is often far from being perfect. We agree with Duckham et al. [8] in that, since imperfection of geographic information is part of the game, it is necessary to develop formal models to measure it.

In this section, we introduce the Geocoding Certainty Indicator (GCI), a method to determine how certain we are as to the correspondence between the given textual address and the resulting coordinates, considering the current contents of the addressing database. The GCI is intended to be used both as a filter, by enabling users to discard whichever data fall below a certain certainty threshold, and as a weighting parameter, allowing statisticians and analysts to use it as a sort of "importance level" associated with the point. It can also be used as a ranking criterion, in the sense employed in information retrieval applications, by which the most probable results are identified, sorted, and presented to the user.

The GCI is a number contained in the interval [0, 1], with 0 meaning completely uncertain and 1 meaning absolutely certain. It is calculated as the product of three sub-indices, all of which ranging in the same interval, namely the Parsing Certainty Indicator (PCI), the Matching Certainty Indicator (MCI), and the Locating Certainty Indicator (LCI). Each of these components will be described next. As an example for the entire process, consider the following string, representing a postal address:

```
RUA FLRIDA, 15—SION—BELO HORIZONTE—MG
```

Notice that `RUA` is the street type, there is a typo in the street name (should be `FLORIDA`), and that `SION` is a neighborhood name in the city of Belo Horizonte, in the Brazilian state of Minas Gerais (`MG`).

5.1 Parsing certainty indicator

The PCI is based on an assessment of how certain can we be as to the correct separation of a textual address into its component attributes, as described in Section 4.1. It is an indication of how "complete" the address is, considering the local addressing system:

depending on the local custom, some parts are more important than others, some parts may never appear at all, and some parts may be indispensable.

Notice that the PCI does not take in consideration the contents of each address component; this is left to the matching phase. For now, it is only important to assess whether the components have been identified or not in the input string. As a result, if the input addresses are already parsed from a structured source, such as a legacy information system, the PCI must be set at 1.

In order to accommodate the great variety of addressing systems, and considering the addressing templates we proposed in Section 4.1, we propose the generation of an 8-bit binary code, in which each bit represents an address attribute (thoroughfare type, thoroughfare name, building number, neighborhood name, building or landmark name, city, state, and postal code). We do not consider the complement, since it usually only indicates units within a building, and thus does not contribute to determine the geographic location. Each bit takes on the value 1 if it has been filled by the parsing routines, and zero otherwise. The resulting set of 256 possible combinations is used as the key column for a look-up table, which receives several values, assigned by the user, each one indicating how close the combination is to a "complete" address in each case, considering a set of local parameters.

As an example, Table 4 shows three combinations and the assigned PCI values, considering generic characteristics of the typical Brazilian, Mexican and Japanese addressing systems. In the first combination, the existence of most addressing elements, except building names, causes the PCI to be high for both Brazil and Mexico. However, building names are important in the Japanese system, therefore the PCI gets a much lower value. In the second line, the Brazilian PCI gets an intermediate value, since there is a city name, but no state name, and there are many coincidental names of cities in different states. The postal code could help in the disambiguation, but it is also not present. This factor is irrelevant for Mexico and Japan, which assign higher values to their PCI. In the third line, the typical Japanese address is followed almost strictly, with no thoroughfare type and no state, thus getting a high value. In Brazil and Mexico, having no state causes a slightly lower value, but the presence of a postal code should allow for the correct identification of the city, even if there are coincidental names. The fourth line shows nearly perfect addresses, which fall short of the 1.00 score just because there is no postal code.

PCI values can either be determined manually, by someone who knows closely the workings of a given addressing system, or determined automatically, by developing formulas in which the absence of certain components diminishes the perceived closeness of each combination to an "ideal" address. These formulas can consider the close match that sometimes is required between address components. For instance, in many places the presence of a postal code can compensate the absence of a city, a state, or even a thoroughfare; in other situations, the postal code provided can be unreliable, and therefore the absence of other components can be considered a great risk, thus requiring a low PCI value.

Table 4 PCI values look-up table

Ttype	Tname	Bnumber	Nname	Bname	City	State	Pcode	BRA	MEX	JAP
1	1	0	1	0	1	1	1	0.85	0.80	0.50
1	1	1	0	1	1	0	0	0.70	0.90	0.80
0	1	1	1	0	1	0	1	0.80	0.90	1.00
1	1	1	1	0	1	1	0	0.95	0.95	0.95

In the example, we obtain the structure presented in Table 5. There is no building name or postal code in the parsed address, and therefore the resulting bit code (11110110) matches the fourth line in Table 4, giving a PCI of 0.95.

5.2 Matching certainty indicator

The matching phase receives the results of the parsing, which, as mentioned earlier, can be incomplete, due to the characteristics of the local addressing system. Only the components received by the matching phase will have an influence on the value of the MCI.

In the matching phase, a set of seven objects forms the ideal result, but only five of those can provide a geographic location (OMU, ONE, OCE, OLM, OIA). If none of them are found, MCI is assigned a value of 0. If OCE, OLM or OIA are found, MCI is assigned an initial value of 1. If only OMU and ONE are found, MCI is assigned a value of 0.5. If only OMU is found, MCI receives a value of 0.25. Furthermore, for each string-matching difference encountered, MCI gets diminished by 0.05, i.e., the MCI is reduced by 0.05 times the edit distance between the original string and the reference string found. Notice that this penalty occurs even in the cases in which there is a unique match, but with an approximation, since such imperfections need to be signaled to the user.

In the example, considering the operations listed in Table 3, we have:

- Operation 1.1 fails, since there is no postal code, and therefore OPC is set to null;
- Operation 2.1 finds multiple matches, since there are three cities in Brazil named “Belo Horizonte”, in different states;
- Operation 2.2 finds the correct city, which is Belo Horizonte in Minas Gerais (MG) state, with an exact string match (edit distance equals zero) and copies the corresponding object to OMU;
- Operation 3.1 finds multiple matches for the neighborhood name, but operation 3.2 finds the correct object, associated to Belo Horizonte, with an exact string match (edit distance equals zero) and then copies the matching neighborhood object to ONE;
- Operations 4.1 and 4.2 are bypassed, since the example address does not have a building name or a complement;
- Operations 5.1 to 5.3 are performed, but generating multiple matches; operation 5.4 succeeds in finding the correct thoroughfare object, with an approximate string match (edit distance is one, since the matching street name is `FLORIDA`) copying it to OTH; operations 5.5 to 5.9 are bypassed;
- Operation 6.1 succeeds in finding the individual address corresponding to number 15 within the thoroughfare, therefore copying the corresponding object into OIA; operations 6.2 to 6.4 are bypassed;
- Operation 7.1 finds a centerline object associated to the OTH object, which includes number 15 in its addressing range, and copies it to OCE. Operation 7.2 is bypassed.

As a result, out of the five location-capable objects, the matching phase has obtained four (OMU, ONE, OCE, and OIA). Therefore, in the example the MCI is assigned a value of

Table 5 Parsed address example

Ttype	Tname	Bnumber	Nname	Bname	City	State	Pcode	Complement
RUA	FLRIDA	15	SION	–	BELO HORIZONTE	MG	–	–

0.95, since OIA and OCE were found (MCI=1), but a typo in the thoroughfare name causes a 0.05 penalty.

The MCI index is calculated making a closed world assumption. Therefore, different address databases may give different values for MCI for the same input address. The different results may provide a parameter that allows users to choose a more appropriate dataset to a certain application. A different use of MCI would be the possibility of using a golden standard for addresses in order to check the accuracy of addressing databases. The MCI then would be a measure of how good is the addressing database.

5.3 Locating certainty indicator

Having received a set of objects from the matching phase, the locating phase proceeds to determine the coordinates from each of them. If this determination can be performed exactly (i.e., by simply copying the coordinates from the object to the result, using the PointCoord function), LCI is assigned the highest possible value, i.e., LCI=1.

If the location is determined along a line, LCI is determined using the following equation:

$$\mathrm{LCI} = 1 - \frac{\mathrm{length(OCE)} \times \mathrm{OTH.Twidth}}{\mathrm{MA}}$$

i.e., the uncertainty is the ratio between the area in which the location can be found (centerline length times thoroughfare average width) and the municipality area (MA). This applies both to the case in which a specific centerline object could be found (method 7.1 in Table 3) and to the case in which range interpolation is not possible (method 7.2 in Table 3).

Likewise, if the coordinates are determined randomly within an area (RandomPoint function), LCI is calculated using the following equation:

$$\mathrm{LCI} = 1 - \frac{\mathrm{area(ONE)}}{\mathrm{MA}}$$

where area(ONE) is the area of the reference object within which the random point has been generated, and MA is the municipality area. If the ONE object is not valid, a random point is generated within the municipality area. If this is the case, then locating fails, and LCI=0.

In our example, LCI is 1, since the location can be determined from the coordinates of OIA.

5.4 Final GCI determination and applications

Finally, GCI can be calculated as the product of PCI, MCI and LCI, as follows:

$$\mathrm{GCI} = \mathrm{PCI} \times \mathrm{MCI} \times \mathrm{LCI}$$

In our example, we have

$$\mathrm{GCI} = 0.95 \times 0.95 \times 1 = 0.9025$$

Notice that the GCI results can be compared to a previously established threshold in order to determine which results are acceptable and which are not. The threshold can be calculated by simulating a hypothetical situation. As an example, suppose a typo is found in the thoroughfare name, another one in the neighborhood name, and a third one occurs in the

city name. These mistakes would bring the MCI down to 0.85; results with a lower GCI can be discarded or reviewed by humans, in order to achieve a greater certainty level.

The GCI has been proposed with the intention of helping the user to pinpoint possible geocoding problems, with a focus on the application. GCI results should provide indications as to address data quality, so that the user can correctly assess the results of a geocoding effort. Statistical analysis of GCI results (as well as analyses on PCI, MCI, and LCI results) can provide the user with indications as to the points to work with in order to improve quality.

Furthermore, information on the quality of the reference database can be obtained as a by-product of the method and of the proposed indicators, if a controlled set of addresses is supplied for geocoding. As an example, consider the situation in which user data are good, but the reference database is outdated. In this case, low GCI results in areas of recent development would be expected. Over large address sets, one might investigate the reason for that by looking for concentrations of low-GCI addresses. If such points are evenly distributed, it would be reasonable to conclude that the problem is mostly with the user source data; if there are concentrations, we might conclude that these areas are outdated in the database, thereby justifying an updating effort.

6 Conclusions

Street addresses are the most common form of linking data in a spatial database to data collected for different purposes such as health, safety, and taxes. Addresses are also an important link to legacy systems that contain valuable data both for historical and updating purposes. Unfortunately the format that real addresses are recorded is not the same as the one used by the spatial database. It is common to have addresses recorded in an unstructured way. Even the addresses that are recorded in a more structured way may have different problems. Therefore the process of matching external addresses to addresses in a spatial database is a complex one. Techniques to improve the matching (called here approximation) and methods to measure the final quality of the geocoding process are needed.

Our first conclusion is that nowadays addresses are important in diverse areas that deal with georeferenced information. Addresses cannot be considered as mere attributes of buildings or of traffic accidents. Addresses are entities in themselves in the modeling level and in the conceptual level. We also described the most common uses of addresses and the different representations that they can take, and gave a formal definition of a general address considering this point of view.

We introduced an indicator of the result of the address geocoding process, the geocoding certainty indicator (GCI). The evaluation of the address certainty indicator took into consideration the spatial transformations that an address record goes through during the matching, and the approximations used to match the external address record with an existing record in the database. The indicator, which has a minimum value of 0 (completely uncertain) and a maximum value of 1 (absolutely certain), is calculated from partial indicators that are determined during the parsing, matching and locating phases. The indicator values, as well as the values of its components, can be used in various situations. We emphasized here the possibility of filtering the data by establishing a GCI threshold, but it is also possible to include the GCI value in the analysis process, as a sort of weighting factor. GCI can also be used as a ranking criterion in geographic information retrieval procedures that use addresses as spatial references.

Since matching addresses from non-spatial sources to spatial entities is an important and useful process, we described algorithms used to do the matching. The quality of the data that are included in the reference database is fundamentally important to the geocoding process. Geocoding over a poor, outdated, or imprecise set of data will produce lower GCI results. On the other hand, GCI results are a function of the current reference database state: improvements in the database can generate higher GCI values from the same input data.

Street addresses are the most common form of linking data in a spatial database to data collected for different information purposes, such as for health, safety, and taxes. It is estimated that 80% of the information used in a local government is associated with addresses. Creating and maintaining an address base is a fundamental step in a successful urban GIS project [9]. In this paper we presented a series of tools that help the automation of the process of locating addresses. We also introduced a measurement to evaluate the certainty of the results of the address geocoding process. Future work includes the development of procedures that allow for the distinction between the uncertainty associated to the input data, and the uncertainty associated with the reference database.

Acknowledgements Frederico Fonseca's work was partially supported by the National Science Foundation under NSF ITR grant number 0219025 and by the generous support of Penn State's College of Information Sciences and Technology. Clodoveu Davis's work is partially supported by CNPq, the Brazilian governmental agency in charge of fostering scientific and technological development. His work in this paper is related to projects ChegoLá (FAPEMIG EDT 1461/03), Saudavel (CNPq grant number 552044/2002-4), and EndFlex (CNPq grant number 502853/2004-2). Authors also thank PRODABEL, the information technology company for the city of Belo Horizonte, for providing data used in the development and testing of the software described in the paper. The authors also wish to thank Max Egenhofer for his comments and suggestions on an early draft of this paper.

References

1. W. Aref and H. Samet. "Optimization strategies for spatial query processing," in *17th International Conference on Very Large Data Bases*, Barcelona, Spain, 1991.
2. K.A.V. Borges, A.H.F. Laender, C.B. Medeiros, A.S. Silva, and C.A. Davis Jr. "The web as a data source for spatial databases," in *V Brazilian Symposium on GeoInformatics (GeoInfo 2003)*, Campos do Jordão (SP), 2003.
3. K.A.V. Borges, C.A. Davis Jr., and A.H.F. Laender. "OMT-G: an object-oriented data model for geographic applications," *GeoInformatica*, Vol. 5(3):221–260, 2001.
4. Britannica Student Encyclopaedia. *Seoul*. Encyclopaedia Britannica Online Volume, 2006.
5. C.A. Davis Jr. "Address base creation using raster–vector integration," in *URISA 1993 Annual Conference*, URISA: Atlanta, Georgia, 1993.
6. C.A. Davis Jr., F. Fonseca, and K.A.V. Borges. "A flexible addressing system for approximate geocoding," in *V Brazilian Symposium on GeoInformatics (GeoInfo 2003)*, Campos do Jordão (SP), 2003.
7. G. Derekenaris, J. Garofalakis, C. Makris, J. Prentzas, S. Sioutas, and A. Tsakalidis. "Integrating GIS, GPS and GSM technologies for the effective management of ambulances," *Computers, Environment and Urban Systems*, Vol. 25(3):267–278, 2001.
8. M. Duckham, K. Mason, J. Stell, and M. Worboys. "A formal approach to imperfection in geographic information," *Computers, Environment and Urban Systems*, Vol. 25(1):89–103, 2001.
9. P. Eichelberger. "The importance of addresses—the locus of GIS," in *URISA 1993 Annual Conference*, URISA: Atlanta, Georgia, 1993.
10. Federal Geographic Data Committee. Draft Proposal for a National Spatial Data Infrastructure Standards Project—Address Content Standard, FGDC, 2003.
11. M. Goodchild. "GIS and transportation: status and challenges," *GeoInformatica*, Vol. 4(2):127–139, 2000.
12. L.L. Hill. "Core elements of digital gazetteers: placenames, categories, and footprints," in *4th European Conference on Research and Advanced Technology for Digital Libraries*, 2000.

13. K. Hiramatsu and T. Ishida. "An augmented web space for digital cities," in *IEEE/IPSJ Symposium on Applications and the Internet (SAINT—01)*, 2001.
14. C.B. Jones, R. Purves, A. Ruas, M. Sanderson, M. Sester, M. Van Kreveld, and R. Weibel. "Spatial information retrieval and geographical ontologies: an overview of the SPIRIT project," in *The 25th Annual International ACM SIGIR Conference on Research and Development on Information Retrieval (SIGIR 2002)*, Tampere, Finland, 2002.
15. P. Longley, M. Goodchild, D. Maguire, and D. Rhind (Eds.), *Geographical Information Systems*. Wiley: New York, 1999.
16. K.S. McCurley. "Geospatial mapping and navigation on the web," in *Tenth International World Wide Web Conference (WWW10)*, ACM: Hong Kong, 2001.
17. A.M.V. Monteiro, M.S. Carvalho, R. Assunção, W. Vieira, P.J. Ribeiro, C.A. Davis Jr., and L. Regis. SAUDAVEL: Bridging the Gap between Research and Services in Public Health Operational Programs by Multi-institutional Networking Development and Use of Spatial Information Technology Innovative Tools. Instituto Nacional de Pesquisas Espaciais: São José dos Campos (SP), Brazil, 2004; Available from: http://www.dpi.inpe.br/saudavel/documentos/ArtigoSAUDAVELAgo2004.pdf.
18. M. Morad. "British Standard 7666 as a framework for geocoding land and property information the UK," *Computers, Environment and Urban Systems*, Vol. 26(5):483–492, 2002.
19. G. Navarro. "A guided tour to approximate string matching," *ACM Computing Surveys*, Vol. 33(1): 31–88, 2001.
20. G.R. Rhind. Global Sourcebook of Address Data Management: A Guide to Address Formats and Data in 194 Countries. Gower: Aldershot, 615, 1999.
21. T.B. Richards, C.M. Croner, G. Rushton, C.K. Brown, and L. Fowler. "Geographic information systems and public health: mapping the future," *Public Health Reports*, Vol. 114(4):359–373, 1999.
22. RuaVista Magazine. The Numbering System of Buildings. [Web site] 2003 [cited 2003 05 may 2003]; Available from: http://www.ruavista.com/numbering.htm.
23. G. Rushton, G. Elmes, and R. McMaster. "Considerations for improving geographic information system research in public health," *URISA Journal*, Vol. 12(2):31–49, 2000.
24. P. Scarponcini. "Generalized model for linear referencing in transportation," *GeoInformatica*, Vol. 6 (1):35–55, 2002.
25. L. Simpson and A. Yu. "Public access to conversion of data between geographies, with multiple look up tables derived from a postal directory," *Computers, Environment and Urban Systems*, Vol. 27(3): 283–307, 2003.
26. L.A. Souza, C.A. Davis Jr., K.A.V. Borges, T.M. Delboni, and A.H.F. Laender. "The role of gazetteers in geographic knowledge discovery on the web," in *3rd Latin American Web Congress*, Buenos Aires, Argentina, 2005.
27. U.S. Census Bureau. 108th CD Census 2000 TIGER/Line Files Technical Documentation. 2003 March 2003 [cited; 321]. Available from: http://www.census.gov/geo/www/tiger/tgrcd108/tgr108cd.pdf.
28. S. Wu and U. Manber. "Fast text searching allowing errors," *Communications of the ACM*, Vol. 35 (10):83–91, 1992.
29. D.H. Yang, L.M. Bilaver, O. Hayes, and R. Goerge. "Improving geocoding practices: evaluation of geocoding tools," *Journal of Medical Systems*, Vol. 28(4):361–370, 2004.
30. J. Zobel and P. Dart. "Finding approximate matches in large lexicons," *Software—Practice and Experience*, Vol. 25(3):331–345, 1995.

Clodoveu Augusto Davis Junior received his B.S. degree in Civil Engineering in 1985 from the Federal University of Minas Gerais (UFMG), Brazil. He obtained M.Sc. and Ph.D. degrees in Computer Science, also from UFMG, in 1992 and 2000, respectively. He led the team that conducted the implementation of GIS technology in the city of Belo Horizonte, Brazil, and coordinated several geographic application development efforts. Currently, he is a professor and researcher at the Pontifical Catholic University of Minas Gerais, and the editor of Informatica Publica, a Brazilian journal on information technology for the public sector. His main research interests include geographic databases, urban GIS, spatial data infrastructures, and multiple representations in GIS.

Frederico Torres Fonseca graduated from the Federal University of Minas Gerais with a degree in Data Processing (1977) and the Catholic University of Minas Gerais with a B.S. in Mechanical Engineering (1978). Subsequently he obtained a Master's in Public Administration from the João Pinheiro Foundation, Brazil (1995). Dr. Fonseca was an Assistant Professor in Computer Science at the Catholic University of Minas Gerais, and held professional appointments as senior systems analyst, GIS analyst and programmer. He completed his Ph.D. with Dr. Egenhofer at the University of Maine in 2001. His thesis covered the area of GIS interoperability. His research provides a theoretical basis for semantic interoperability, which is a necessary foundation for a working, interoperating environment. With his focus on system design, Dr. Fonseca demonstrates how complicated processes can be integrated to the benefit of users. His newly developed concept of ontology-based GIS is highly interdisciplinary as it brings together various research methods from artificial intelligence, software engineering, and GIS. Dr. Fonseca is the recipient of the 1999 ESRI/IGIF Scholarship, of the 2000 Graduate Research Assistant Award and of a NASA/EPSCoR fellowship. Currently, he is an Assistant Professor at the College of Information Sciences and Technology at the Pennsylvania State University.