Geocoding: Fundamentals, Techniques, Commercial and Open Services

Franz-Josef Behr

Department of Geomatics, Computer Science and Mathematics, University of Applied Sciences Stuttgart
Schellingstraße 24, D-70174 Stuttgart, Germany

KEYWORDS:

Spatial Data Infrastructure, Interoperability, Web services, Address data, Geocoding, Open Geodata

ABSTRACT:

Referencing information and features to geographic locations is an essential component of daily live. In newspapers and television news we are accustomed to see maps illustrating the location of events in politics, sports, or natural hazards. The process of assigning location to such features is called georeferencing.

Postal addresses, a specific way of locational description, are the essential means by which people express their location in the real world. Also many features in modern societies in administration and business are related to addresses which are used as unique identifiers to geographical locations, often expressed by a pair of longitude and latitude values. Although addresses are data containing location, they do not contain coordinates. Therefore an important feature in many applications especially in Geographic Information Systems (GIS) is the capability to locate addresses, i. e. to geocode to address level.

During the last few years there has been also a significant progress to geocode addresses by commercial Internet mapping application programming interfaces (APIs) and Internet services. Recent advances, mainly in the field of Internet technology, have encouraged people to develop and to integrate addressing data models from variety of addressing system. It has made geocoding to be more popular, and using such online tools, the geocoded address can be retrieved and displayed. The process of geocoding is not perceived as an individual service by the user but as a standard function of WebMapping.

In the beginning of this paper, fundamentals of address schemas and of geocoding approaches are introduced. The process of parsing addresses and matching with reference datasets is explained. Phonetic algorithms including Soundex and Kölner Phonetics, Levenshtein distance are explained in this context.

Generally, geocoding can be performed effectively on the basis of point locations of buildings. If such data is not available, the position can be derived by interpolation along street axises as available in navigation data. These methods are discussed including their pros and cons.

Some open solutions available on the web are reviewed and compared concerning their coverage, the APIs offered, and the response formats:

The paper concludes by reviewing two open, participatory, community oriented projects: one is an open geocoding service (opengeocoding.org), the other an open geodata project: (openaddresses.org). Both projects rely on data derived from volunteer work.

1 Introduction

1.1 Motivation

One particular data set for a spatial information infrastructure is address data, the base of many administrative, statistical and economic processes in every country (Lind 2010). There are a variety of applications relying on address information like marketing, planning, emergency services, route planning, crime mapping, environmental health studies etc. During the last few years there has been also a significant progress to geocode addresses by commercial Internet mapping application programming interfaces (APIs) and Internet services, mainly based on street network data collected for navigation systems. The development of technology, mainly in the field of Internet technology, has encouraged people to develop and to integrate addressing data models from variety of addressing system. It has made geocoding to be more popular and online mapping tools have popularized the concept of using an address as an initial map navigation tool. Using such online tools, the geocoded address can be retrieved and displayed. The process of geocoding is not perceived as an individual service by the user but as initial step for web mapping (Gruber 2006).

While in few countries such data is already provided by state agencies or offered by commercial vendors, in most of the developing countries such data is not available at all (Vivas 2010).

Based on such street data and address ranges, online mapping tools offer since a few years geocoding services, but limited to developed countries. Today, when the Internet has revolutionized the world of Information Technology, users from developing countries should not be left behind: Even in less developed countries addresses should be available as well as a geocoding application to provide locational information (i.e. coordinates) for postal addresses. Such a geocoding application will not only integrate open data with open source code, but will have also its own database to support the demand of users in developed and developing countries.

1.2 Background

1.2.1 Geospatial features and footprints

Georeferencing deals with geospatial features. Such features can be natural physical features, e.g., lake, rivers, or mountains, but geospatial features can be also man-made, physically existing like cities, roads, buildings, or virtual features, like postal code areas, regions, provinces, and countries, or statistical data.

The (geo-) referencing of geospatial features by placenames is, as Hill (2006) denotes, the most commonly used form of referencing a geographic location in human communication. Such placenames (in a strict way we should say: feature names) are used in many circumstances, in daily conversation, titles of books and articles, but also in digital information items, like news feeds, catalogue entries, HTML pages, or messages.

A placename (toponym) can be a single term (the proper name), but in many cases it is extended by additional modifiers (Hill 2006):

- Administrative modifier: Administrative modifiers (like county, state, country) setup a hierarchy of administrative units for different administrative levels. Examples for such modifiers are "Washington D.C." or "Monte Carlo, Principauté de Monaco". In some case of ambiguities a city can be identified by administrative modifiers like county or state.
- Locational modifier: A locational modifier helps to identify a certain location by adding an additional placename. This type of modifier is often used in conjunction with city names, like "Stratford-upon-Avon" or "Rothenburg ob der Tauber".
- Place type or functional modifier. Toponyms are often additionally labeled by a generic part indicating the type of location, e.g., "Lake Constance" or "Mount Titlis", "New York City". Quite often however, these terms are implicitly assumed (e.g., "Bavaria" instead of "State of Bavaria", or "New York, New York").

For many purposes geospatial features are categorized according to specific fields of application. Some generic categories can be mentioned.

- Topography, representing the terrain, lakes, rivers, etc.;
- Planimetry, like parcels, buildings, and other man-made features;
- Roads, comprising different road types;
- Thematic data, like environmental measurements, or social data.

In some application fields, like in water or electricity supply, we can even find hundreds of different categories.

Geospatial features have a geospatial footprint, e.g., coordinates, which can be categorized (following Open Geospatial Consortium 2006) as Point, LineString, LinearRing, Polygon, and collections of these primitives (MultiPoint, MultiLineString, MultiPolygon, MultiGeometry). The geospatial footprint becomes especially important if maps or computerized representations of geospatial come into play.

Because the footprint of geospatial features can be very detailed and complex, it is often simplified to a generalized form, e. g., a point such as a centroid with latitude and longitude value, or a minimum bounded rectangle (MBR) as it is provided in the GML standard instead of the real geometry. Such representatives can be efficiently stored and evaluated for information retrieval even if the DBMS used is not capable of spatial operations (Hill 2006:88). In addition, they are more easily to obtain, nowadays even for layman using Global Positional System (GPS) sensors. The coordinates are related to a spatial reference system. In most cases we use two-dimensional coordinates, in a generic way called easting and northing. In many cases, especially in context with Internet based applications and the GPS geographical coordinates, e.g., longitude and latitude values are used, in form of decimal degrees. Besides this notation, longitude and latitude values are sometimes traditionally noted in the form of degrees, minutes, seconds.

1.2.2 Gazetteers

Collections of placenames are called *gazetteers*. Typically each entry consists of three parts: the placename, its type / category, and its geospatial footprint. Gazetteers are used to provide the translation (i.e. geocoding) of placenames to spatial footprints and for the reverse process, finding placenames based on coordinates (i.e. reverse geocoding). Gazetteers are set up for many purposes, on local, regional, and world-wide level. Geocoding of placenames can use one or more gazetteers if corresponding interfaces to access their data are offered.

1.2.3 Representing address data

Postal addresses are one often used way of specifying locations. The INSPIRE Directive (INSPIRE 2007) defines this kind of the spatial data theme Addresses as the "Location of properties based on address identifiers, usually by road name, house number, postal code."

Several approaches exist for the representation of postal address data:

- Point locations: Geocoding relying on individual point locations yields the highest accuracy. In several countries datasets are already offered by public agencies or private companies, in most cases available for high fees only. Addresses can be directly matched to the addresses of the reference data set.
- Street segment interpolation: If no individual point locations are available, the street network of a municipality can modeled based on the center lines for named streets. Each line segment has a start and an end node providing a sense of direction for a named street. Address ranges are assigned for each side (left and right) of the street segment. Typically, it is assumed the even and odd house numbers are on different sides of the street. For many developed countries, such data is already available.
- Addresses firstly have to be matched with the street names of the reference data set. If the match is successful, the house number is interpolated spatially according to address ranges applying the even/odd assumption (cf. **Fehler! Verweisquelle konnte nicht gefunden werden.**). Additionally a mostly fixed offset from the street center line is added. It is evident that the interpolation and the fixed offset can have significant impact on the accuracy of the resulting coordinates. In reality, houses can be spaced unequally, and the offset can vary heavily (cf. **Fehler! Verweisquelle konnte nicht gefunden werden.**). Additional errors can be introduced by missing street segments or overlapping streets (Ahlers & Boll 2009).
- Areal features: Areal features, like countries, states, provinces, or postal code areas are represented by a point. Address geocoding is based solely on this information.

1.3 Related Applications

Based on these approaches, a large number of geocoding and reverse-geocoding applications are available (http://groups.google.com/group/Google-Maps-API/web/resources-non-google-geocoders). However, by analyzing such an enumeration it can be noted that most of theses services are offered mainly for industrialized countries like USA, Canada, Japan, and the EU states.

Many Web applications, so-called mash-ups, have been implemented based on commercial geocoders. Majewski (2006) explains the geocoding features of a widely used geocoder, Google's Geocoding Application Programming Interface (API). Addresses can be submitted as a single string and are parsed on server side, expanding abbreviations. The geocoder response uses JSON (see Wagner et al. 2009), KML or xAL format (explained below). Generally, developers are encouraged to build their own client-side caches. Because geocoding via HTTP is supported, it is possible to use AJAX technology (Garrett 2005) on client-side or to call the service from server side applications.

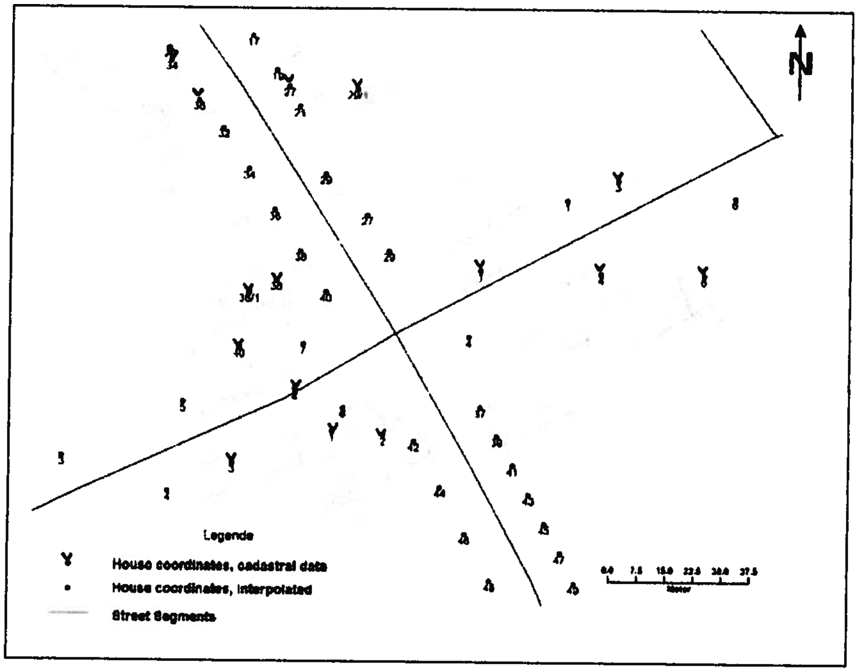

N
Legende
Y House coordinates, cadastral data
• House coordinates, interpolated
Street Segments
0.0 7.5 15.0 22.5 30.0 37.5
Meter

Figure 1: True address location, based on cadastral data, and interpolated address locations (Source: Schäfer & Kunz 2008, reproduced by permission of the authors).

The Geocoding API of Yahoo! Maps supports the Representational State Transfer approach (REST, Richardson & Ruby 2007). The result is delivered in XML format. Similar to Google this service also provides a precision response about the result of the geocoding process. Theurer (2005), for example, describes a mash-up using Yahoo!Maps which offers geocoding addresses, and visualization with additional information layers The mash-up uses REST to call the Yahoo! API.

Both services have severe limitations concerning the frequent usage. Yahoo's geocoding service is limited to 5,000 queries per IP address per day; Google's service is limited to 50,000 requests per day per API key which is related to a specific domain name.

TeleAtlas, a leading data provider for many commercial mapping services offers its own geocoding service named EZ-Locate (http://www.geocode.com/). This service provides fee-based access to the Tele Atlas address database and geocoding technology using web browser, API based requests, or a stand-alone PC-based client.

All these approaches are characterized by their binding to commercial vendors and the data sets offered. In contrary, there are also freely available services. Geonames.org offers a GUI based geocoding and mapping service for country and place names as well as postal codes based on freely available data. Geocoding is also supported by REST web services. Since April 2008, the originally free service has a 50'000 credits limit per day and IP address. For each geocoding service, 1 – 4 credits are needed.

Geocoder.us (http://geocoder.us/) is basically a Perl frontend to the TIGER/Line files from the US Census Bureau to geocode US addresses or street intersections. Here it is evident that the data are based on governmental street data provided for the whole country.

Universal Address System, developed by NAC Geographic Products Inc. (http://www.nacgeo.com/), is designed to provide a highly efficient unified representation of an address, postal code, area code, geographic coordinates and map grid coordinates. A geocoding service is offered as well.

There are also several free projects of geocoding applications. Geokit (http://geokit.rubyforge.org/) is a Ruby on Rails plugin for mapping applications geocoding, location finders, and distance calculation. For geocoding Geokit provides a uniform interface for other geocoders (Google, Yahoo, Geocoder.us,), and IP-based geocoding. Another geocoding library for Ruby as a wrapper for Yahoo!Maps and Geocoder.us is available from http://geocoder.rubyforge.org/. Two further, community oriented application are described in more detail in the sub-sections below.

1.4 Geocoding Accuracy

Geographic locations are generally available with a limited accuracy and, in some cases, with some vagueness related to the place name (e.g., Southern France, Bavarian Alps). Such uncertainty can be dealt by the fuzzy-set approach, which is out of scope of this paper. Nevertheless it has to be noticed that the limited accuracy has impacts for the results of a geocoding process and all further analysis steps (Grubesic & Murray 2004). The final accuracy is influenced by the accuracy and errors in the reference data and by the errors introduced by the geocoding processes due to approximation and interpolation.

Several researchers present results of accuracy assessment. Ratcliffe (2001) compared TIGER-type geocoded address data in relation to cadastral and census areal units. As a result from a study of over 20000 addresses in Sydney, Australia, he detects that 5–7.5% are misallocated to census tracts, and that more than 50% may be geocoded within a different land parcel.

Davis et al. (2003) discussed how geocoding can be implemented over incomplete and possibly inaccurate addressing data (which is the prevailing situation in many municipalities in developing countries) and how users can benefit from the results of such a process.

In the context of public health research Cayo & Talbot (2003) determined positional errors for 3,000 residential addresses using the distance between each geocoded point and its true location as determined with aerial imagery. Error was found to increase as population density decreased. The mean error for rural areas was 614 m, 143 m for suburban areas and 58 m in the urban area.

Nicoara (2005) investigated the lack of standardization in addressing, differing ways of defining location, ownership and property definitions, the complexities of multi-source location and data address maintenance, differing address table schemas and models, and the variety of algorithms used. Borges et al. (2007) examined the automatic process of recognition, extracting, and geocoding of addresses in Web pages. They report successful geocoding (for city granularity) up to 77%.

Stark (2008) reports an accuracy assessment of the geocoding engines of Google Maps and Microsoft Virtual Earth. 90,000 addresses of the Canton of Solothurn were used to test the geocoding accuracy; official cadastral data – assumed as true location – were used as reference. The Google Maps geocoder returned 93% of the addresses with an accuracy level of 6 (street level accuracy). Only 53% of these addresses have a distance less or equal to 20m from the true location. With Microsoft Virtual Earth, 86% of the addresses were classified as a good match; in this case, only 41% show a distance lower or equal to 20m compared to the reference data.

Ahlers & Boll (2009) report their approach of combination of two of the most widely used geocoders for obtaining optimal results. Strength of single services can be used, while weaknesses can be avoided.

1.5 Address Standards

If placenames consist of postal addresses, the diversity of address formats come into play. Interoperability problems have encouraged many organizations and developers to create specifications for name and address data (Cover 2008). The Organization for the Advancement of Structured Information Standards (OASIS, http://www.oasis.org) has quite successfully developed the extensible Customer Information Quality Specifications Version 3.0 (CIG) which includes formal definitions for Name (xNL), Address (xAL), Name and Address (xNAL) and Party (xPIL). One of them, xAL, is used as part of Google Gecoder's response. Based on xNL and xAL specifications, xAL uses hierarchical nested XML elements to encode address information and is also available as one of OpenGeocoding's output options (section 3.6). Besides these general address formats, local developments are found. The Government of Victoria (2006), for example, published its own address standard.

Also international boards are involved in case of standardization. Schulzrinne (2006) proposed a Request for Comments to IETF. Information about the country, administrative units such as states, provinces, and cities, as well as street addresses, postal community names, and building information is suggested as part of the Dynamic Host Configuration Protocol. The option allows multiple renditions of the same address in different scripts and languages. Issues related to an international address standard, its relevance for postal and Spatial Data Infrastructures and other topics were part of ISO/TC 211's agenda during a workshop in Copenhagen end of May 2008 (http://www.isotc211.org/Address/Copenhagen_Address_Workshop/index.htm).

Addresses are also treated in INSPIRE Directive Annex I. INSPIRE (2010) describes the INSPIRE Data Specification on Addresses

Addresses and geocoding are also part of the standardization work of the Open Geospatial Consortium (OGC). Its OpenGIS Location Services (OpenLS) document[1] describes an abstract, standardized set of interfaces, protocols and data types for implementations of Location Based Services (LBS) including geocoding and reverse geocoding.

2 The Geocoding Process

The process of geocoding generally includes several steps:

1. Parsing and structuring the address / locational description;
2. Matching process, including quality assessment;
3. Proliferation of the results.

A typical input for this process is a semi-structured addresses that contains single items for address components such as street, postcode, and city name.

A formal concept of postal addresses is given by Borges et al. (2007). The basic address supplies a street type, its name, and a house number. The second, optional part provides a complement to the basic address, including neighborhood name. The third part, called location, consists of three identifiers applicable in urban context: postal code, phone number, and/or city/state. An address can be complete (having all necessary parts), incomplete (having the basic part plus one of the identifiers), or partial (consisting only of one of the three identifiers. It is worth noting that there is some redundancy in postal addresses. Hence the postal code can give additional support if a city name is ambiguous.

A similar concept is given by INSPIRE (2010:7). The address must be associated with a number of "address components" that define its location within a certain geographic area. Each of the address components represents a spatial identifier as for example the name of a road, district, postcode, municipality, region or country. Remarkable, an address has besides the geographic position which enables an application to locate the address spatially a "locator", e.g. a unique identifier (e.g., address number and / or name) distinguishing it from the neighbour addresses and temporal attributes.

Language and cultural specific ordering of the parts has to be taken into account while parsing an address ("Geo-parsing") to convert this semi-structured address to a structured address according to the structure required by the matching process.

According to Mikheev et al. (1999, cited by Clough 2005) simple list lookup for place names already can perform well, but during the matching process some obstacles may occur (Ding et al. 2000):

- Alteration of placenames over the time: For example "Saint Petersburg" formerly was called "Leningrad" and "Petrograd". People might use different names for the same location.
- Aliasing: Different names are commonly used for the same location on a regional level and in different languages as well. For example "München" is called "Munich" in English, and "Monaco" in Italian.
- Ambiguity: A placename refers to locations in different countries. For example "Saint Petersburg" exists in Russia, and in the USA in Florida, and Pennsylvania. Such a placename without further qualifier is inherently ambiguous; a ranking can be done according to the probability that a certain location is the right one.
- Vernacular naming: "Hyères" in France is called "Iero" or "Ieras", depending on the Provençal dialect, spoken in the south of France.
- Multilingual naming: Address matching becomes even more difficult in regions that are bi- or multilingual. I.e. in Switzerland there are cities where street names exist both in German and French. Additionally the order of addressing parts are altered according to the language used, e.g., while the German version of an address is street name, house number. the French way is house number, followed by street name.
- Spelling errors: Spelling may not be equal to the one in the reference data-set. Missing accents have to be considered as well as diacritical marks,
- Omissions: Quite often in daily practice, placename modifiers are omitted. Such anticipations can cause additional efforts if a geocoding application has to support different languages.

[1] http://www.opengeospatial.org/standards/ols

Such particularities and ambiguities need to be taken into consideration when geocoding addresses by applying specific string matching algorithms. Sometimes spelling errors or different variations in spelling of an address may occur and the geocoding mechanism must deal with the problem and decide whether the textual location description can be matched to the reference data-set at all or whether some minor or major changes in the spelling need to take place. For the matching entries in the gazetteer, the matching quality can be expressed, for example as an indicator from 0 to 100.

Geocoding addresses is generally not perfectly successful due to spelling differences or errors either in the reference data-set or in the address itself. Unmatched addresses or matches on a lower level of granularity furthermore decrease the accuracy of the matching process, e.g., if resolution does not go down to building level but only to street, postal code or even place-name or country level,

Further challenges in address-parsing and validating are discussed in Ahlers & Boll (2008).

2.1 Granularity, Accuracy, and Precision

While assessing the results of a geocoding process, it is necessary to distinguish several, related terms which are used varyingly by different authors and services.

Generally, the geocoding can be achieved on different levels of spatial granularity. It can be performed on base of a street address, or on a lower level of granularity, like city or country. The number of levels of granularity can be different, depending on the geocoding engine used. Ahlers & Boll (2008) introduce five classes in terms of spatial granularity (country, region, city, street, building). Google's geocoder report one of 10 different levels of granularity[2] (0 = unknown location, 1 = country level, 2 = region (state, province, prefecture, etc.), 3 = sub-region (county, municipality, etc.), 4 = town (city, village), 5 = postal code (zip code), 6 = street, 7 = intersection, 8 = address, 9 = premise (building name, property name, shopping center, etc.)). Yahoo instead uses 7 level of granularity[3].

Accuracy indicates the level to which the resulting coordinates adhere to the real location. In most instances, the geocoding engine can not report this value because the accuracy can not predicted in many cases. However, as illustrated in section 1.4, several authors examined the accuracy of geocoding services and obtained results which depend on service, area, and other parameters.

Precision, finally, refers to the exactness of the digital representation of the resulting coordinates, often recognizable by the number of digits. As Hill (2006) emphasizes "that precision does not imply accuracy in any way".

2.2 Placename Matching

String matching algorithms have been discussed by many authors. Exemplarily White (2005) gives a good overview of different matching approaches. Basically there are two main categories of matching algorithm for string similarity: the first class comprises equivalence methods and the second similarity ranking methods.

Equivalence methods compare two strings and return a value: true or false. The decision is made according to whether the method judges two strings to be equivalent. Word stemming is one approach which tries to find a common stem of two terms. Also synonyms are used to find perfect matches. With addresses especially abbreviations are used to say the same thing: blvd for boulevard, st for street, ave for avenue, rd for road and others.

Similarity ranking methods on the other hand compare a given string to a set of strings ranking these reference strings in terms of similarity. A numeric measure indicates the measure of similarity for each comparison.

Levenshtein distance[4], for example, is defined as the minimal number of edit operations necessary to convert one string into the other. Edit operations can either be insertions, deletions or substitutions of characters. Cologne phonetics is a phonetic string matching algorithm, originally published by Postel (1969), and finally typewriter distance takes the position of keys on the typewriter-keyboard and their relative alignment to each other into consideration.

[2] called by this service "accuracy", see http://code.google.com/intl/en/apis/maps/documentation/reference.html#GGeoAddressAccuracy. The term accuracy here indicates however only the level of granularity, not a metrical accuracy.

[3] called by this service context "precision" http://developer.yahoo.com/maps/rest/V1/geocode.html

[4] http://www.levenshtein.de/

Another approach for matching algorithms is phonetic string matching algorithms. Zobel and Dart (1996) introduce the use and mode of operation of phonetic string matching algorithms. With these algorithms spelling is used to identify other strings that presumably have the same pronunciation. Implicitly this is very much language-dependant. Soundex (Knuth 1998) is one of the most popular representatives of this class (Zobel & Dart 1996, Zandbergen 2007).

Some of the algorithms mentioned above, especially the Soundex algorithm, are implemented in Database Management Systems like MySQL®, PostGreSQL and Oracle® and programming languages.

3 Open Address and Geocoding Services

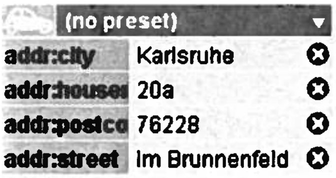
3.1 Overview

The idea of Open Address and Geocoding services is simple and follows the approach of volunteered geographic information, well described by Goodchild (2008) as Volunteered Geographic Information (VGI), one of the characteristics of the Web 2.0. Every person has a local spatial expertise and can act as a "living sensor". If this local knowledge is globally collected in a structured way a valuable and comprehensive information repository is established for a multitude of use cases. Therefore the challenge is to connect these "sensors", collect and organize their expertise and finally provide this accumulated information for public usage, as described by Amelunxen (2010).

3.2 OpenStreetMap.org

The OpenStreetMap (OSM) project currently is one of the most successful VGI projects. Up to now approximately 1.8 billion GPS points have been uploaded by more than 250.000 registered users, converted to 700 Million nodes (points) and 56 Million ways (linear features)[5].

Users can tag the data entered by key value pairs in a very flexible way, e.g. key=amenity, value=Parking or key=Highway, Value=Residential. In addition they are uniquely identified by an identification number. Following these rules, an address schema was established owning its own addr namespace comprising the keys housenumber, housename, street, postcode, city, country and additional keys for interpolation along ways (kind of interpolation, e.g. all, even, odd, alphabetically, number increment used for interpolation, and a tag to indicate the accuracy level of survey used to create the address interpolation, e.g. actual/estimate/potential)[6].



addr:city	Karlsruhe	⊗
addr:house	20a	⊗
addr:postco	76228	⊗
addr:street	Im Brunnenfeld	⊗

Figure 2: Example of tagging a postal address in OpenStreetMap.

Based on these proposed tags users collected more than 4 Million nodes tagged with values for city, country etc. and more than 5 Million with values for street name[7]. Address input is done using one of the editors like Potlatch, JOSM, or Merkaartor without special interface for address input.

For Geocoding several approaches are supported by OpenStreetMap[8].

Nominatim is the current search system for OpenStreetMap. In addition to a GUI-based research (from the OpenStreetMap website) REST-based requests are supported offering interoperability with other web services. As result formats, HTML, XML, and JSON can be requested.

Name Finder (http://wiki.openstreetmap.org/wiki/Name_finder) offers search for place names, postal addresses, and supports a "near" search. Name Finder was historically used for geocoding based on the OSM database.

[5] http://www.openstreetmap.org/stats/data_stats.html [2010-07-21]

[6] http://wiki.openstreetmap.org/wiki/Addresses [2010-07-21]

[7] http://tagwatch.stoecker.eu/Planet-100707/En/grouplist.html [2010-07-21]

[8] In addition The geocoding of Geonames.org is integrated in the OpenStreetMap portal

Currently the Name Finder database is dated "9 Jan 2009"[9] and seems out of synchronization with the central OSM database.

3.3 OpenAddresses.org

Unlike other volunteered approaches for geodata capturing, OpenAddresses.org does not necessarily require data-capture with a GPS sensor in the field. Data collection is primarily organized through a mapping mash-up, currently based on OpenLayers[10]. The user interface is currently available in several languages (e.g. German, English, French, Italian, Polish, etc.). The collected data is provided under GNU General Public License version 3.

Figure 3 shows how the user is required to interactively and manually digitize the locations of the corresponding addresses using the map data layers presented in the mash-up (OpenStreetMap and Yahoo imagery). Stark (2009) presents the predecessor project whole project in more detail.

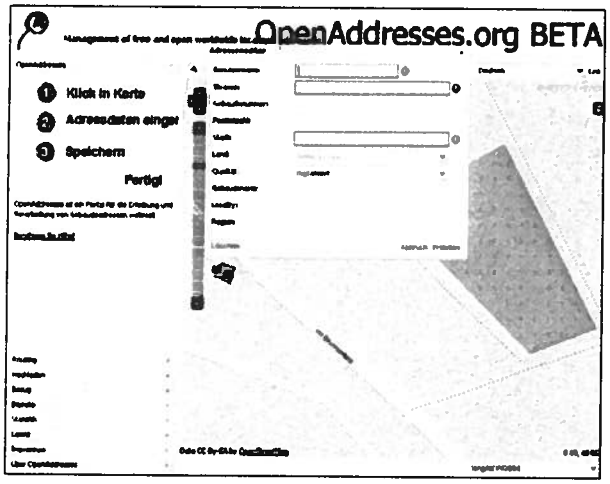


Figure 3: The OpenAddresses.org website offers an interactive geocoding function.

REST-based requests are supported offering interoperability with other web services. As result formats, HTML, XML, and JSON can be requested.

3.4 OpenGeocoding.org

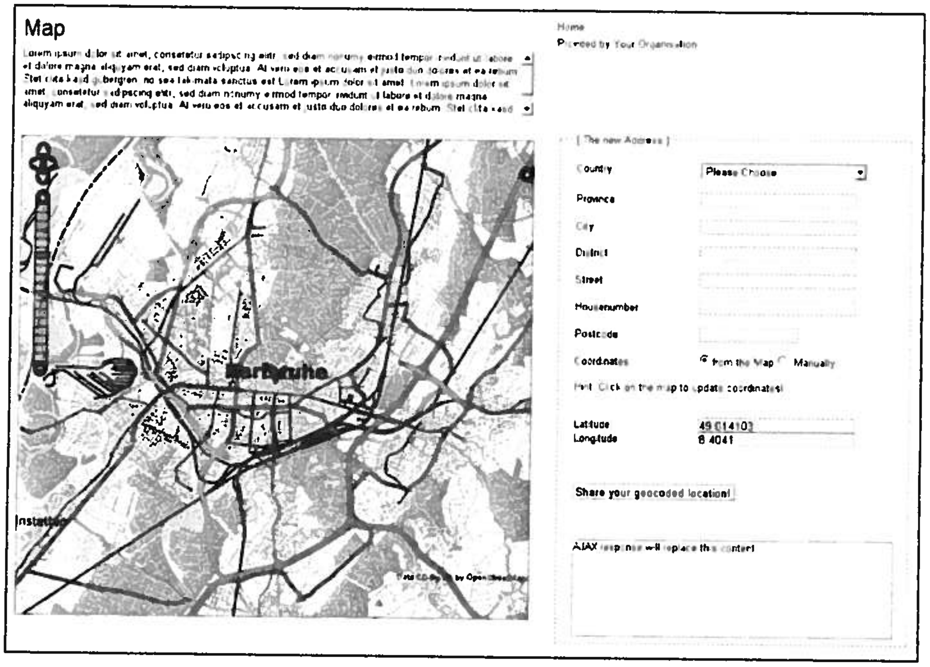
While in developed countries usually governmental or private nation-wide efforts to generate and maintain address data exist, in less developed countries such geodata are still not available (Vivas 2010). The main objective of this project is the generation of a worldwide free address database for developed *and* developing countries – the application aims to collect data as well to provide a geocoding service.

- Data collection: Registered users submit address information and corresponding coordinates (figure 4). Special attention is given to the validation to prevent spamming, by comparison of entered data to existing reference data on country, province and city level. Supported by AJAX based techniques (Garret 2005) for map display and user friendly auto-completion, users can enter addresses and the corresponding coordinates which could be derived from maps or GPS measurements. Alternatively, users can use a Web map (currently with several data layers, e. g., OpenStreetMap and different Google® Maps layers) to digitize the location. After successful data validation, address and coordinates are saved in the database and can be used for geocoding, interactively or through the API.
- Geocoding: The geocoding service for postal addresses is offered based on this data free of charge through a REST based web service (Richardson & Ruby 2007). The geocoding application does not rely only to its internal database filled by the community. If no relevant results can be retrieved, the request will be forwarded to other publicly available geocoding services. The result will be checked, parsed, and forwarded to the user in a specified format.

[9] http://gazetteer.openstreetmap.org/namefinder/ [2010-07-21]

[10] http://openlayers.org/

Besides offering REST services, the application additionally offers a Web based user-friendly front-end for interactive address submission.



Powered by Your Organisation

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd

[The new Address]

Country	Please Choose
Province	
City	
District	
Street	
Housenumber	
Postcode	
Coordinates	● from the Map ○ Manually

Hint: Click on the map to update coordinates!

Latitude	49.014103
Longitude	8.4041

AJAX response will replace this content

Figure 4: Geocoding using OpenGeocoding.org.

3.4.1 Technologies used

The portal site is based on the Openlayers API, JavaScript and AJAX. PHP is the corresponding server-side scripting language for database access, geocoding and creation of web pages. MySQL, since February 2008 owned by SUN Microsystems, Inc. (http://www.mysql.com/) is used as a database management system.

3.4.2 Output formats

The service provides results in XML, KML, GML, JSON, and CSV format (see Wagner et al. 2009). The XML format corresponds to the aforementioned xAL standard, embedded in a KML envelop generated according to the OpenGIS® KML Encoding Standard (http://www.opengeospatial.org/standards/kml/). GML output contains coordinates according to the OpenGIS Geography Markup Language (GML) Encoding Standard (http://portal.opengeospatial.org/files/?artifact_id=20509).

4 Conclusion

The "worldwide" geocoding problem is far from being solved for a variety of reasons. From the technical point of view it involves the development of standards, availability of reliable data in a digital format and associated issue of accuracy. In many countries, the main problem is the availability of data (especially in a digital format) since many municipalities often do not have street names! The applications presented here are an approach to overcome these restrictions by participatory, community based data capture. These applications can be an alternative to the existing geocoding services, independent from the commercial solutions provided that a critical mass of participants can be achieved.

References

Ahlers D., Boll S. 2008: Retrieving Address-based Locations from the Web. Proc. Second International Workshop on Geographic Inforamtion Retrieval, GIR'08, October 29–30, 2008, Napa Valley, California, USA.

Ahlers D., Boll S. 2009: On the Accuracy of Online Geocoders. Proc. GEOINFORMATIK 2009, Osnabrück, Amelunxen C. 2010: An Approach to Geocoding based on Volunteered Spatial Data. http://www.geoinformatik2010.de/public/abstracts/amelunxen.pdf [2010-07-19]

Borges. K. A. V., Laender, A. H. F., Medeiros, C. B., Clodoveu A. D. Jr. 2007: Discovering Geographic Locations in Web Pages Using Urban Addresses. GIR'07, November 9, 2007, Lisbon, Portugal

Cayo M. R., Talbot T. O. (2003): Positional error in automated geocoding of residential addresses. http://www.ij-healthgeographics.com/content/2/1/10 [2009-05-27]

Davis Jr., C. A., Fonseca, F. T., De Vasconcelos Borges, K. A. (2003). A Flexible Addressing System for Approximate Geocoding. Proceedings Geoinfo 2003. Available online at http://www.geoinfo.info/geoinfo2003/papers/geoinfo2003-25.pdf. [2009-05-27]

Davis Jr., C. A., Emerson de Salles E. (2007): Approximate String Matching for Geographic Names and Personal Names. IX Brazilian Symposium on GeoInformatics, Campos do Jordão, Brazil, November 25-28, 2007, INPE, p. 49-60.

Cover, R, (2008): Markup Languages for Names and Addresses. http://xml.coverpages.org/namesAndAddresses.html [2009-05-27]

Clodoveu A. Davis Jr., Emerson de Salles (2007): Approximate String Matching for Geographic Names and Personal Names. IX Brazilian Symposium on GeoInformatics, Campos do Jordão, Brazil, November 25-28, 2007, INPE, p. 49-60. http://www.geoinfo.info/geoinfo2007/papers/S3P2.pdf [2009-06-08]

Clough, P. 2005. Extracting metadata for spatially-aware information retrieval on the internet. In Proceedings of the 2005 Workshop on Geographic information Retrieval (Bremen, Germany, November 04 - 04, 2005). GIR '05. ACM, New York, NY, 25-30. http://ir.shef.ac.uk/cloughie/papers/cikm2005-geo.pdf

Ding, J., Gravano, L., Shivakumar, N. 2000. Computing Geographical Scopes of Web Resources. In *Proceedings of the 26th international Conference on Very Large Data Bases* (September 10 - 14, 2000). A. E. Abbadi, M. L. Brodie, S. Chakravarthy, U. Dayal, N. Kamel, G. Schlageter, and K. Whang, Eds. Very Large Data Bases. Morgan Kaufmann Publishers, San Francisco, CA, 545-556.

.Garrett, J. J. (2005). AJAX: A New Approach to Web Applications. Technical report, Adaptive Path. http://www.adaptivepath.com/publications/essays/archives/000385.php. [2009-05-27]

Goodchild M. F. (2008): Volunteered geographic Information, GEOconnexion International Magazin, October, 2008

Government of Victoria, 2006. Whole of Victorian Government Standard - Street Address Data Standard. A standard for the storage, interchange and validation of street address data by Victorian Government agencies. http://www.dtf.vic.gov.au/CA25713E0002EF43/WebObj/-DIStreetAddressData/$File/DI%20Street%20Address%20Data.pdf [2009-05-27]

Grubesic T. H., Murray A. T. 2004: Assessing positional uncertainty in Geocoded Data. Proc. Urban Data Management Symposium, 2004

Gruber, F. 2006. Comparing the Mapping Services. http://www.techcrunch.com/2006/04/17/comparing-the-mapping-services [2008-05-09]

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

INSPIRE 2007: Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE) 14.03.2007. http://eur-lex.europa.eu/JOHtml.do?uri=OJ:L:2007:108:SOM:EN:HTML [2010-07-19]

INSPIRE 2009: Draft COMMISSION REGULATION implementing Directive 2007/2/EC of the European Parliament and of the Council as regards interoperability of spatial data sets and services 14.12.2009. http://ec.europa.eu/transparency/regcomitology/-index.cfm?do=Search.getPDF&lA6b4z6edALEzOuvQ2DQwEuYwr24bl+u6M8oCwqlYrvB7EJR+poTzWZ/2wT/z/JFTr7x0HnynbCJdi/BzR4ZvdPpAur0FOHhej8jYcN49FA=^ [2010-07-19]

INSPIRE 2010: INSPIRE Data Specifications on Addresses - Guidelines v 3.0.1 03.05.2010. http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_AD_v3.0.1.pdf [2010-07-19]

Knuth. D. E. 1998: The Art of Computer Programming, Vol. 3: Sorting and Searching.2. Aufl., Reading/Mass.: Addison-Wesley.

Levenshtein, V. I. (1965). "Binary Codes Capable of Correcting Spurious Insertions and Deletions of Ones." Probl. Inf. Transmission 1: 8-17.

Lind, M. 2010: Benefits for society of a strong address infrastructure - Danish case study. European Address Forum Brussels, 15 June 2010, http://www.europeanaddressforum.eu/-EAF/index.php?option=com_attachments&task=download&id=39 [2010-07-19]

Mikheev A., Moens M. and Grover C. (1999) Named Entity recognition without gazetteers. In Proceedings of the Annual Meeting of the European Association for Computational Linguistics EACL'99, Bergen, Norway, 1-8.

Nicoara, G., 2005. Exploring the Geocoding Process: A Municipal Case Study using Crime Data. http://charlotte.utdallas.edu/mgis/prj_mstrs/2005/Summer/greta/Nicoara_Masters/Website/main.html.[2008-05-09]

Open Geospatial Consortium 2006: OpenGIS Implementation Specification for Geographic information - Simple feature access - Part 1: Common architecture http://portal.opengeospatial.org/files/?artifact_id=18241 [accessed 2009-05-22]

Postel, H. J. (1969). Die Kölner Phonetik. Ein Verfahren zur Identifizierung von Personennamen auf der Grundlage der Gestaltanalyse. IBM-Nachrichten. 19, 925-931.

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, Vol. 15, no. 5, 473 – 485

Richardson L., Ruby S. (2007): *RESTful Web Services*. O'Reilly

Stark H.-J. (2008): Open Geodata - am Beispiel von OpenAddresses.ch, in: Strobl, J. et al. (Hrsg.), Angewandte Geoninformatik 2008. Beiträge zum 20. AGIT-Symposium Salzburg. Heidelberg, 2008

Stark H.-J. (2009) OpenAddresses - Free geocoded street addresses. Proc. Applied Geoinformatics for Society and Environment (AGSE) 2009, http://applied-geoinformatics.org/ [2009-06-14]

Schäfer B., Kunz C. 2008: Geocoding. Unpublished project Stuttgart University of Applied Sciences

Theurer, D. 2007: How to build a Maps Mash-up. http://www.theurer.cc/blog/2005/11/03/how-to-build-a-maps-mash-up/ [2010-07-20]

Vivas, P. 2010: "Addressing the world – An address for everyone". European Address Forum Brussels, 15 June 2010, http://www.europeanaddressforum.eu/-EAF/index.php?option=com_attachments&task=download&id=35

Wagner, D., Zlotnikova, R., Behr. F.-J. 2009: XML-Based and Other Georelated Encodings: Overview of Main Existing Geocoding Formats. In: Behr, F.-J., Schröder, D., Pradeepkumar, A. P. (Editors) 2009: Proceedings Applied Geoinformatics for Society and Environment (AGSE 2009). Publications of the Stuttgart University of Applied Sciences, Hochschule für Technik Stuttgart, Volume 103 (2009), 317 pp., ISBN 978-3-940670-13-7, http://www.gis-news.de/papers/AGSE_Proceedings_2009_07_01_1025.pdf [2010-07-20]

White S. (2005): Tame the Beast by Matching Similar Strings. http://www.catalysoft.com/-articles/MatchingSimilarStrings.html?article=Tame_the_Beast_by_Matching_Similar_Strings_14 [2009-05-27]

Zandbergen P. A. 2007: A comparison of address point, parcel and street geocoding techniques. Computers, Environment and Urban Systems, Volume 32, Issue 3, May 2008, Pages 214-232

Zobel J., Dart P. (1996): Phonetic String Matching: Lessons from Information Retrieval. www.cs.rmit.edu.au/~jz/fulltext/sigir96.pdf [2009-05-27]