Multi-Source Toponym Data Integration and Mediation for a Meta-Gazetteer Service

Philip D. Smart, Christopher B. Jones, Florian A. Twaroch

School of Computer Science, Cardiff University
{c.b.jones,p.d.smart,f.a.twaroch}@cs.cf.ac.uk

Abstract. A variety of gazetteers exist based on administrative or user contributed data. Each of these data sources has benefits for particular geographical analysis and information retrieval tasks but none is a one fit all solution. We present a mediation framework to access and integrate distributed gazetteer resources to build a meta-gazetteer that generates augmented versions of place name information. The approach combines different aspects of place name data from multiple gazetteer sources that refer to the same geographic place and employs several similarity metrics to identify equivalent toponyms.

1 Introduction

Automated gazetteer services that maintain geo-data associated with geographic place names are becoming increasingly important for a variety of applications [1]. They are needed to recognise place names that users employ in queries to retrieve geographic information and for applications that need to detect the presence of place names in text resources, for example to index documents for a spatially-aware search engine. Gazetteer services are also required for the reverse geocoding process of finding place names associated with geographical coordinates, e.g. to attach a place name to a GPS-referenced photo. To provide effective support for these sorts of applications raises the challenge of creating a gazetteer that can maintain access to a wide range of place name terminology relating to many different sorts of features at arbitrary locations. In practice however, because the quality of the content of a gazetteer will depend upon the application for which it is required [2], it is not possible to specify the characteristics of a single ideal gazetteer, except perhaps at a generic level.

At present the number and types of gazetteer resources are increasing as commercial gazetteers become supplemented by volunteered sources of geographic place name data. The gazetteer sources differ considerably with regard to their geographical coverage, the range of features types, the presence of alternative names, and the detail and accuracy of their geometric footprints. National mapping agencies (NMA) generate gazetteers that relate typically to their respective geographical areas, and while they may be reasonably reliable with regard to representation of formal administrative geographic entities, their range of types of named feature and of the terminology employed is inevitably limited to particular, usually topographic, themes. Commercial sources of place name knowledge may have wider geographical extents

but are usually focused on particular types of application, especially that of navigation, and tend to reflect the administrative view of geography. Volunteered place name resources, such as the Geonames gazetteer and OpenStreetMap may support a range of features types and levels of detail not found within the commercially-marketed and NMA sources. It is also the case that gazetteer sources differ in their data structures and access methods. Some need to be loaded to a database, while others have web service or other interfaces.

Given these varied sources of gazetteer knowledge, there is a motivation to create a "meta-gazetteer" that accesses multiple resources in order to retrieve the best toponym information available for a particular purpose. Because the different sources of place name information differ in the quality of their associated data, such as the feature types and the spatial footprints, there is need for a form of conflation [3] in which multiple sources are compared and merged so that the best aspects of each source can be combined. In the present paper we address these requirements and describe methods for multi-source place name data integration and mediation that have been developed to support a distributed web gazetteer service, that is used for geo-parsing free text and for reverse geo-coding. After a review of related work in section 2 we give an overview of our Toponym Ontology data model and the associated mediation-based architecture (section 3). Section 4 analyses the characteristics of the formal and volunteered sources employed and section 5 explains our mediation system methods for selecting, integrating and augmenting toponym geofeatures from multiple resources. We present results and an evaluation of the geofeature matching procedure in section 6 and give then an outlook to future work.

2 Related Work

Typical data that are stored in gazetteers are standard and alternative names of a geofeature, the type of the named feature, a geometric coordinate-based footprint, such as a point, a bounding box or a polygon, and one or more parent features within an administrative or topographic hierarchy [1]. Gazetteer specifications such as that of the Alexandria Digital Library support further attributes such as spatial and other relations between features, the data accuracy, and the source of the data for an individual place.

In a recent review of requirements for a next generation gazetteer, Keßler et al [4] drew attention to a number of desirable gazetteer properties which include those of accessing multiple data sources, exploiting volunteered data sources, maintaining mechanisms to assess trust in resources, development of an agreed high level domain ontology, inclusion of deductive inference of knowledge and the development of a semantically enabled user interface.

Gazetteer enrichment from web resources still faces a number of challenges: for example, place name recognition methods applied to text corpora degrade significantly in their performance (measured in recall and precision of developed GIR systems) when no gazetteer is initially considered [5], i.e. to construct a gazetteer we need a gazetteer. Uryupina [6] presented a method to detect place names together with their feature type using a search engine and machine learning techniques but does not

focus on a particular geographic region. Natural language processing methods to detect place names in text corpora do not ground them with geometric footprints.

Goldberg et al [2] created enriched name and feature type data for merged address (parcel) level places using multiple representations of the same place derived from online residential and commercial phone books. Equivalence of features was established in terms of equivalence of the name and address attributes. This simple testing was facilitated by a prior data cleansing or normalisation process that transformed their sources to a common USPS address format, using the probabilistic "record linkage" methods of Christen and Churches [7]. A third "official" county web site data source was used to validate derived addresses.

Flickr tags have been used to build representations of place [8, 9]. While this is a step in the right direction the approaches depend heavily on the availability of volunteered geographic information in a single source. Data integration for gazetteer construction has been addressed by Hastings [3], though not in the context of a distributed access environment. His conflation methods employ geotaxical and geonomical semantic similarity metrics. Gazetiki constructs a gazetteer that integrates geographic concepts found on Wikipedia pages with the location derived from Panoramio photos for several European cities [10].

3 Overview of Toponym Ontology Model and Mediation System

We introduce a toponym ontology (TO), which is equivalent to what others may call a gazetteer, in combination with web service and data mediation functionality that enables access to multiple resources in response to a query on the TO. The main purpose of the TO is to support geo-parsing and reverse-geocoding (see section 1). Thus the TO has to 1) find toponym-data that matches a given input string representing a place name and 2) retrieve georeferenced place names given a spatial footprint as an input. For task 1 accurate coordinate data and wide geographic coverage are required from the gazetteer resource while for task 2 rich hierarchical information is required to provide unambiguous multi-part (hierarchical) toponyms.

The TO model, illustrated in Fig. 1, is based on the concept of a geofeature that corresponds to a named, spatially focused, geographical phenomenon, and conforms to Goodchild and Hill's definition of a place [1]. As with many gazetteers these minimum requirements are supplemented by other components of information. In particular this includes data about the source of the toponym data, the language of the name itself and of its alternative names, and hierarchical links to parent geofeatures of which the current geofeature is a part. The footprint may take the form of a point, a line, a simple polygon or a minimum bounding box, as well as a collection of one or more of these simple types. The footprint is also associated with the definition of the spatial reference system and a datum if available. The feature types are taken from a concept ontology developed in parallel with the TO and which includes scene types that were developed to support the specific applications of the TO.

The TO is to a large extent a virtual store of geofeatures. A query to the TO results in retrieval of TO data from remote geo-data sources, which are integrated on the fly. Because some of the toponym resources are not supported by web service access, the

remote access procedures are supplemented by local database storage of these resources. For reasons of efficiency the TO maintains a local cache of the results of remote access calls and of the results of the integration and augmentation procedures. The framework for remote access is based on the mediation architecture introduced by Wiederhold [11]. It resembles the distributed retrieval engine approach of Callan [12] in being able to access, format and integrate local or remote geo-data sources on demand.

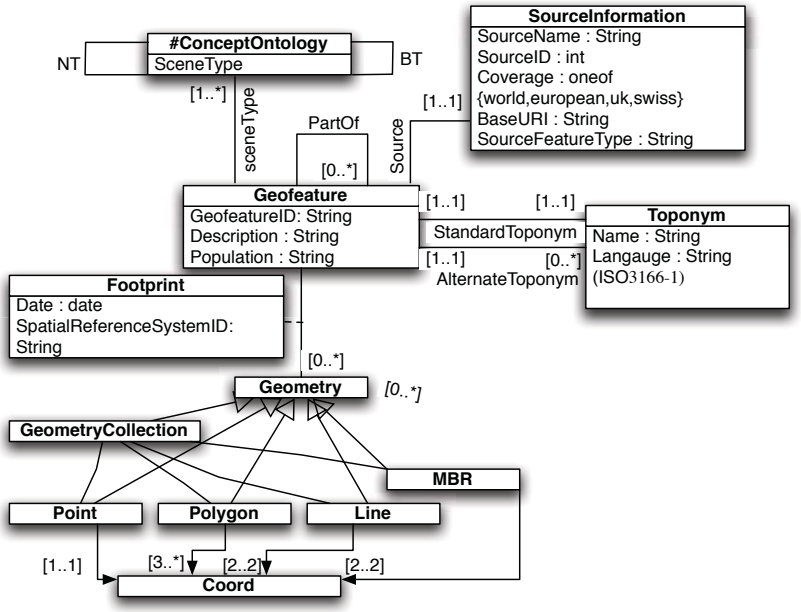


Fig. 1. Toponym Ontology Model

A three layer mediation architecture [11, 13] is employed consisting of a Foundation layer, Mediation layer and Application layer (see Fig. 2). The resource manager acts as an intelligent mediator handling on the fly access to the multiple heterogeneous toponym resources. The remote sources (e.g. Geonames[1]) are connected to and queried via their web service endpoints (Foundation Layer). Local resources are maintained in spatial databases and queries are issued using database connectors. The principal components of the resource manager are the interfaces to each resource (termed *Interface* in Fig.2) , the *Geofeature Integration Module* (GIM) and the *Geofeature Augmenter*. We summarise the characteristics of resource interfaces in this section, while deferring explanation of the GIM and Geofeature Augmenter until section 5.

Each of the accessible data resources employs its own data schema and output data formats, while each of the remote sources also employs its own set of web services adapted to their respective schema. To deal with this the Resource Engine implements

[1] http://www.geonames.org/

a separate interface to each resource. These interfaces query the resource's end point (local or remote) and formats results according to the internal geofeature model of the Toponym Ontology. The Resource Interface is comprised of two components, the *query translator* and the *results translator*.

In view of the divergent data schema and web service interfaces, the interface and internal data model of the toponym ontology can be both logically and syntactically incompatible with each resource [14]. The purpose of the *query translator* is therefore to morph standardised geofeature queries, issued to the Toponym Ontology engine, to a form that correctly queries each resource (where the schema of each resource can easily be interrogated [15]), and hence returns the information required to instantiate one or more geofeatures. Consequently, the resource manager is a fat mediator, in that all processing of source information is performed internally, and is not delegated to each source [16].

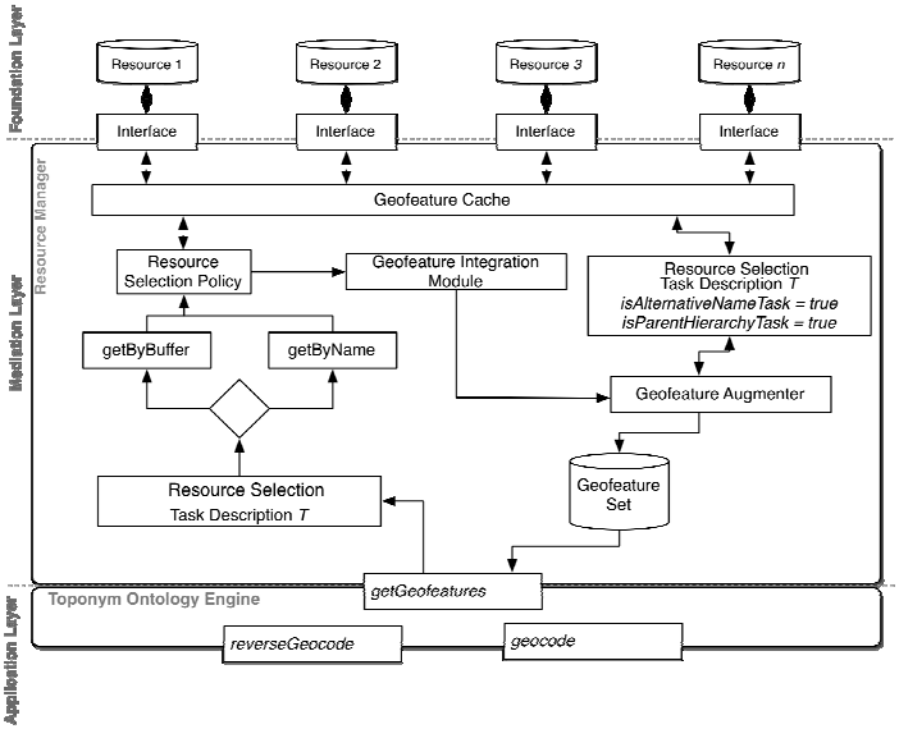


Fig. 2. The mediation architecture

The purpose of the *results translator* is to convert between the native result format of the resource and the Toponym Ontology geofeature model. Translation is important in order to maintain a uniform internal data model, and is defined manually per resource as part of the resource interfaces (Fig. 2).

4 Analysis of Data Resources Accessed by the Toponym Ontology

4.1. Data Sources

At present the remote sources employed are Geonames gazetteer, OpenStreetMap[2] (OSM), Yahoo Where on Earth[3] and Wikipedia[4] georeferenced articles, all subject to Creative Commons Licences. These are either accessed on demand or periodically downloaded in bulk. Local stored data, with no web access methods, include the Ordnance Survey 1:50,000 scale gazetteer (OS50K) and parts of their PointX and Mastermap products. Table 1 summarizes the characteristics of the resources.

Yahoo Where on Earth (YahooWOE), attempts to provide a permanent unique identifier (a WOEID) for every place on the Earth's surface. The standard API does not provide any reverse geocoding functions. Instead, reverse geocoding functions are obtained from Flickr, which has an extended API that wraps the existing YahooWOE API.

Table 1. Resources of typed geofeatures used in the topoym ontology : (H) indicates that a hierarchy is supported.

Source	*# Features*	*Cov.*	*Geometry*	*Access*	*Format*
Geonames	*7 Mio. (H)*	*World*	*Point*	*Remote / Local*	*XML*
OSM	*Unknown*	*World*	*Point/ Polygon*	*Local*	*ESRI shape*
YahooWOE	*Unknown (H)*	*World*	*Point*	*Remote*	*XML*
Wikipedia	*~ 12. Mio*	*World*	*Point*	*Local*	*XML, RDF*
OS Point X	*3.9 Mio*	*GB*	*Point*	*Local*	*GML*
OS 50K	*~ 260k*	*GB*	*Point*	*Local*	*Ascii*
OS MasterMap	*> 10 Mio.*	*GB*	*Point/Polygon*	*Local*	*GML*

Many of the 12 million plus articles in the online multi-lingual collaborative encyclopaedia Wikipedia, have geographic content, being georeferenced by a latitude and longitude, and including alternative names. We imported a database dump of georeferenced articles into the TO. The remaining three Ordnance Survey data sets in table 1 are commercial administrative-oriented products.

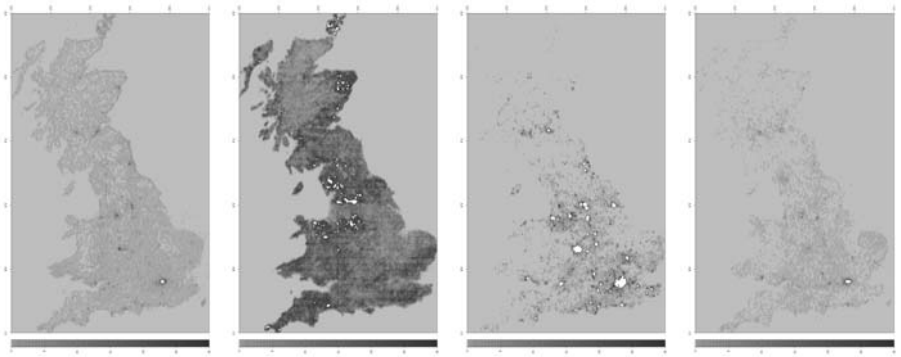
4.2. Comparison of sources

Here we compare locally stored version of four resources – Geonames, Wikipedia, OSM and OS50K - with regard to spatial distribution and some aspects of their associated attribute data. YahooWOE is excluded, due to being remote access only, as are Mastermap and PointX due to limited availability in our project. Spatial distribution is analysed using quadrat counts [17] that divide an area into rectangular sub regions of equal size. The number of toponyms whose footprint intersects each of these quadrats is then counted and recorded for each quadrat (see Fig. 3 and Table 2).

[2] http://www.openstreetmap.org/
[3] http://developer.yahoo.com/geo/
[4] http://www.wikipedia.org/

With regard to spatial distribution of toponyms the OS50K has the most uniform distribution across urban and rural areas. It has also the highest average number of toponyms per quadrat and the highest total number of toponyms. For certain areas however each of the other sources has higher local maximum numbers of toponyms per quadrat than OS50K, particularly within cities. For the UK, Geonames has the second best uniformity of spatial distribution with very good uniformity of distribution for Europe as a whole. OSM is notable for a bias towards to urban areas.

With regard to types of features in the resources, OSM, has relatively few larger scale features such as towns, cities, mountains and countries, but it records many buildings and some landmarks. However there is a bias to commercial service locations such as places of entertainment, ATM machines, garages and restaurants. This emphasis upon small scale features makes it applicable to reverse-geocoding applications for finding small scale features, such as within a photograph's viewport, as opposed to geocoding applications which require data sources rich in prominent landmarks along with larger scale features such as neighbourhoods, towns, cities and countries.



(a) (b) (c) (d)

Fig. 3. 2D filled contour plot of the quadrat count for each of the four data sources in Great Britain, where each quadrat is 3.3km wide and 5.6km high. The contour plot for each source has the same range of z values and hence colour gradient. Values above 30 are coloured white. (a) Geonames, (b) OS50k, (c) OSM and (d) Wikipedia

OS50K has many settlements and a range of other types of features but their feature typing suffers from a very large number of "other" unclassified features (about 50% for example in the city of Edinburgh and more in some rural areas). It is also limited by the low resolution of its coordinates (+/- 500m). Geonames is particularly rich in town/city scale features as well as natural geographic features such as rivers, lakes, mountains, coasts and valleys. Wikipedia is notable for having the highest numbers of well-known landmarks, which makes it good for reverse geocoding applications. Settlement classification for Wikipedia has limitations – for example many relatively small settlements are classed as "city". Both Geonames and YahooWOE have consistent, but different, parent hierarchies.

Table 2. Number of toponyms per web resource

	Geonames	*Wikipedia*	*OSM*	*OS50K*
Total Toponyms	31674	13030	49119	255182
Avg. Toponyms per Quad	0.79185	0.32575	1.227	6.379
Max Toponyms per Quad	245	300	998	58

Geonames hierarchies are administrative and can only be determined for Geonames toponyms, while YahooWOE provides smaller scale neighbourhood (sometimes vernacular) levels up to country level parents with Global coverage. It should also be pointed out that Geonames and Wikipedia are good sources of alternative names which are useful for recognizing places for geocoding purposes. Geonames provides both language variations, alternative spellings and, on occasion, vernacular names.

6 Data Matching and Augmentation Procedures

Here we describe the functions of the resource manager for registering resources, and for matching, integrating and augmenting geofeatures from those resources. The resource manager registers all accessible resources, by adding entries into the resource register, including the name, ID, spatial coverage, data license, the uniquely constructed resource interface and the suitability of each resource. The suitability is encoded as Boolean flags associated with each of a set of tasks for which the Toponym Ontology may be employed. These include geocoding, reverse geocoding, toponym hierarchy construction and alternative names retrieval (see Table 3).

Table 3. Suitability of toponym resources for tasks

	Geonames	*Wikipedia*	*OS50k / OSM / Point X*	*YahooWOE*
Geocoding	√	√	√/-/-	-
Rev. geocoding	√	√	-/√/√	-
Hierarchy	√	-	-	√
Alternative names	√	√	-	-

6.1 Geofeature Integration Module (GIM)

The Geofeature Integration Module (GIM) matches equivalent geofeatures from heterogeneous data sources. Equality is measured on the standard name and spatial location, using a spatiotextual similarity measure. The GIM operates over all sets of Geofeatures G (where $G = \{g_1,\ldots,g_n\}$, and g is an individual geofeature) returned from each source after a get geofeatures request has been issued.

Textual similarity between the standard names is determined using a combination of the Levenshtein metric to measure edit-distance [18], text normalisation (using ICU4J decomposition) and the SoundEx phonetic algorithm [19]. Text normalization uses IBMs International Components for Unicode (ICU) Java library (ICU4J[5]), which

[5] http://www.icu-project.org/

transforms composite characters into pre-composed characters. For example Zürich becomes Zu{dieresis}rich, where the diacritic mark {dieresis} has been removed from the character glyph (u with an umlaut). Removal of all decomposed diacritical Unicode characters results in a canonical form i.e. Zurich. The Levenshtein metric $Sim_{lvd}(w_1,w_2)$ measures the edit-distance between two strings w_1 and w_2, which is the number of edits (alterations such as copy, delete, insert, substitute) needed to change one string to another. Each type of edit is assigned a weighting. If the edit-distance is > 3 its score is set to 0 (edit distances > 3 indicates the two strings are too dissimilar to be considered), otherwise the score is computed as:

$$Sim_{lvd}(w_1,w_2) = \begin{cases} 0 & \text{if } lvd(w_1,w_2) > 3 \\ 1 - \dfrac{lvd(w_1,w_2)}{3} & \text{otherwise} \end{cases} \quad (\mathbf{1})$$

where, $lvd(w_1,w_2)$ computes the edit distance between two strings, and the final similarity measure is in [0,1], where 1 represents equivalent strings.

The SoundEx algorithm matches phonetically similar sounding words, using language dependent rules that allocate numerical values to phonetically distinct character groups. The SoundEx similarity measure computes the difference between two word strings w_1 and w_2. $sdx(w_1,w_2)$ will be a score from 0-4 where 0 represents no similarity and 4 indicates the strings are identical. The final measure $Sim_{sdx}(w_1,w_2)$ is a value in [0,1] where 1 denotes two strings are identical:

$$Sim_{sdx}(w_1,w_2) = \frac{sdx(w_1,w_2)}{4} \quad (\mathbf{2})$$

The combined edit distance and Soundex distance measure (denoted *sim*) is:

$$sim(w_1,w_2) = \frac{(4 * Sim_{sdx}(w_1,w_2) + 1 * Sim_{lvd}(w_1,w_2))}{5} \quad (\mathbf{3})$$

The weighting here is the result of parameter tuning during empirical testing, and reflects the higher confidence we have in using Soundex to match misspelt words. Note that toponyms are always normalised before $sim(w_1,w_2)$ is calculated.

The distance in metres between each pair of geofeatures $<g_1,g_2>$ that have a standard name similarity score > 0.8 is calculated using the geodesic arc distance [20] based on coordinates in the WGS84 spheroid coordinate system that is used for each source (following transformation from their original coordinate system as necessary). Each name-matched pair of geofeatures $<g_1,g_2>$ is then treated as a match if d is less than some value *min* (currently set at 50 metres).

The resulting set G may contain many matched pairs and, for each pair, only one geofeature; g_1 or g_2 is returned. Which to remove depends on the priority of the source of g_1 compared to the priority of the source of g_2. The source with the lower priority is removed from the pair. The list (manually created) of source priorities starting with the highest is: Wikipedia, PointX, Mastermap, Geonames, OSM, OS50K. The removal of matching geofeatures is an iterative procedure in which a lower matching priority geofeature will be removed from all pairs in which it occurs.

6.2 Resource Selection

The Resource Manager has two types of resource selection policies: 1) *Priority Selection* queries each source in a defined order until the query can be satisfied. 2) *Maximum Selection* queries each source in turn and filters the results, using the toponym matching procedure described above. The selected representation may then be augmented with data from other representations of the same place in a later processing stage described in the next section. *Priority Selection* is used by the geocoding function of the toponym ontology for which a single geofeature match is appropriate. Reverse geocoding uses the *Maximum Selection* policy to obtain as many geofeatures as possible within a given spatial buffer. The function *isSuitableFor* takes as input the current resource *s,* and an input task description *T*, and returns true if the resource is suitable for the current task. A task description *T* indicates whether the task is geocoding, reverse geocoding, hierarchy retrieval, alternative names retrieval, or it can specify a particular source.

Algorithms for the two policies are presented in pseudo-code below where $Q(s_i)$ performs a query consisting of either *getByBuffer* (which returns all features within a given buffer) or *getByName* (which returns a set of geofeatures based on a fuzzy standard and alternative name match) on the given resource s_i. Source priority is currently, starting with highest priority: Geonames, OS50K, Wikipedia, Mastermap, OSM and PointX. The procedure *GIM(G)* employs the duplicate detection method described in the previous section. The resource manager can accept other types of query on the registered resources, e.g. relating to retrieval of hierarchical levels.

Algorithm: Priority Selection	**Algorithm: Maximum Selection**
Input: a geofeature query *Q,* a task description *T;* **Output:** a collection of geofeatures *G,* or an empty set if none was found; Let *S* be the set of registered sources; Order *S* by source priority; For each source s_i in *S* If *isSuitableFor*(s_i, *T)* Let $gc = Q(s_i)$; If $gc \neq \theta$ Return *gc;* Return. θ	**Input:** a geofeature query *Q,* a task description *T;* **Output:** a collection of geofeatures *G;* Let *S* be the set of sources in the resource Registry; Let *G* be an empty set of geofeatures; For each source s_i in *S* If *isSuitableFor*(s_i, *T)* Let $gc = Q(s_i)$; Let $G = G \cup gc$ Remove duplicates $G = GIM(G)$) Return *G*

6.3 Geofeature Augmentation

The resource manager can not only retrieve but also construct augmented geofeatures. Information from a number of sources is merged, on the fly during each request, to

create more complete and consistent instantiations of geofeatures. The *Geofeature Augmenter (GA)* performs 1) *addition* of a consistent and accurate set of administrative parents to small scale geofeatures, e.g. POI; and 2) *reconstruction* of full and consistent geofeature records given an arbitrary geofeature.

We present a procedure for administrative hierarchy augmentation. It uses the YahooWOE hierarchy data to augment geofeatures retrieved from data sources such as OSM which have no explicit parent hierarchy, or OS50K and Wikipedia which only contain county or country level parents.

For administrative regions the algorithm only assigns a country level parent, as they already represent part of the administrative hierarchy. Large scale geofeatures such as lakes, parks and others are only given country level parent information, as they could span a number of administrative areas.

Algorithm: Administrative Parent Hierarchy Augmentation

Input: A geofeature *g*

Output: The same geofeature *g* with enhanced parent hierarchy from YahooWOE

Create task description *T*, set *sourceName* = YahooWOE
Let *P* be the set of parents returned by querying getParentHierarchy(*g.location,T*)
Attach the parent hierarchy *P* to *g*
Return *g*

The algorithm *Geofeature Reconstruction* presented below takes as input a geofeature *g* and outputs an augmented version of the same geofeature, following a process of replacement or addition of attributes from matching geofeatures. In the procedure documented here the task description simply specifies a single resource of Geonames that is to be used for matching against the input, as this resource is known to be good quality with regard for example to alternative names, population statistics and feature type. It is possible to extend this procedure to augment from multiple sources. The reconstruction procedure uses a function *STEquiv* which takes as input a set of geofeatures retrieved by a query to the resource and matches them to the single input geofeature, using the GIM matching methods. If an equivalent Geonames geofeature is not found, *g* is returned unchanged.

Algorithm: Geofeature Reconstruction

Input: A geofeature *g*

Output: The same geofeature *g* with enhanced attributes

Create task description *T*, set *sourceName* = Geonames
Let *G* be the set of geofeatures returned by querying
    getByName(*g.standardName,T*)
Let *ge* = *STEquiv (g,G)*
If *ge* $\neq$ null
        Set null values of *g* to those of *ge*
Return *g*

As an example of geofeature reconstruction, consider the following. YahooWOE is a good resource for finding parent containment for point locations through the Flickr API[6]. However YahooWOE is often limited with respect to the number and types of attributes returned. Fig. 4 illustrates augmentation for the geofeature Cardiff, which was retrieved from YahooWOE as the 'Region' parent of the location 51.47, -3.19. The place record retrieved from YahooWOE for Cardiff is typical in having no alternative names, no population information and only a rather broad type classification i.e. 'Region'. Consequently, this record is augmented with alternative names, population information and a more appropriate place type from a matching Geonames entry.

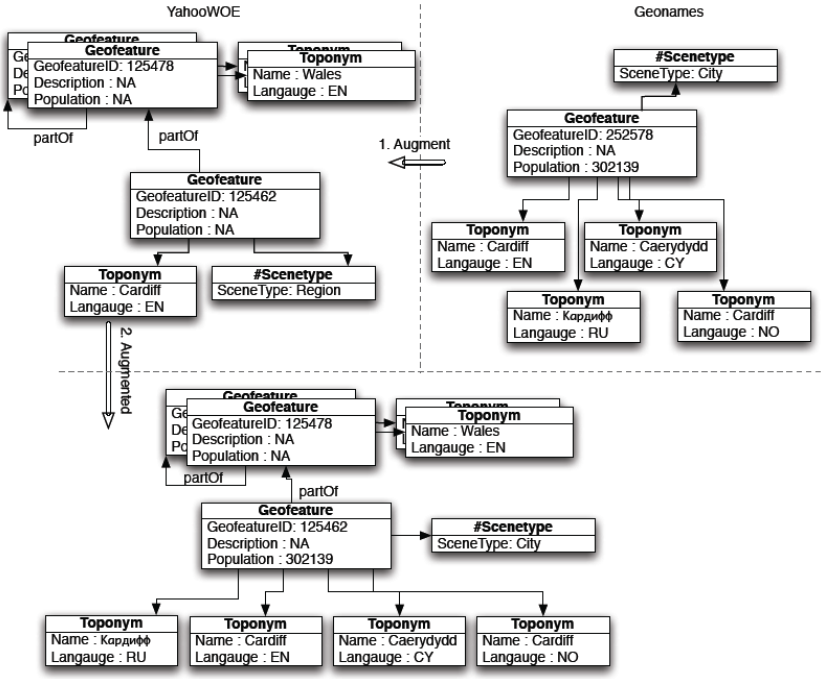


Fig. 4. Example reconstruction of the Cardiff Yahoo WOE geofeature with the Cardiff Geonames geofeature.

7 Results and Evaluation

To test the accuracy of the Geofeature Integration Module (GIM), five different locations were sent to the *getByBuffer* method of the resource manager (each with a 200m buffer), and comparisons from the GIM over the set of returned geofeatures were manually examined. Table 4 shows each of the five locations, the number of

[6] http://www.flickr.com

geofeatures returned from the resource manager, the number of manually identified matching pairs, the number of matching pairs which were successfully matched and resolved (where one is then removed), the number of similar pairs which were not matched (failures), and those that where matched but should not have been (false positives).

The results show the *GIM* to be successful in removing 67.50% of matching names between sources. It is also notable that, for this sample set, the spatio-textual measure does not produce any false positives, i.e. there are no generated matches which are known, by manual investigation, to be incorrect.

Observations of successful matches include: high accuracy in matching locations which start with similar word-grams and have similar locations e.g. 'Cardiff Arms Park (Cardiff RFC)' (OSM) and 'Cardiff Arms Park' (Wikipedia) which have a textual

Table 4. Evaluation of the *getByBuffer* method

Location	*Returned Geofeatures*	*Similar Pairs*	*Success*	*Failures*	*False Positives*
1 – Cardiff	5	2	*1 (50%)*	*1 (50%)*	0
2 – Edinburgh	5	2	1 (50%)	*1 (50%)*	0
3 – Cardiff	2	1	1 (100%)	0 (0%)	0
4 – Edinburgh	*9*	3	3 (100%)	0 (0%)	*0*
5 – Cardiff	*25*	8	3 (37.5%)	5 (62.5%)	*0*
Average			67.5%	32.5 %	0%

similarity score of 0.85 and a distance of 41m; identical name matches between sources e.g. 'Royal Botanic Garden Edinburgh' in both Wikipedia and OSM with a 48m difference between locations; and subtle differences in punctuation e.g. 'St James Centre' in OSM compared to 'St. James Centre' from Wikipedia.

Observations of failed matches include: one source having duplicate entries with initial word-gram name variations that give low Soundex similarity scores e.g. 'Cardiff Millennium Stadium' (Geonames) and 'Millennium Stadium' (Geonames); use of abbreviations in sources leading to large word variations and high edit-distance scores e.g. the user contributed entry in OSM 'Univ Liby' compared to its proper name 'University Library' from Mastermap; and locations with high textual similarity but separated by distances exceeding 50m, e.g. 'Pont Sticill' (Geonames) and 'Pontsticill' (OS Gazetter) with 0.933 name similarity but a 745m difference in location.

8 Conclusions

This paper has addressed the need to access heterogeneous gazetteer data available in the combination of volunteered and formal resources. Our mediation-based meta-gazetteer service supports integration methods that conflate multiple attributes from the different resources using a toponym feature matching procedure. The resource selection and priority strategies were based on a prior analysis of their data characteristics in combination with application requirements. Future work will present the results of already conducted application-oriented evaluations that demonstrate the effectiveness of the methods presented here in practice. It will also focus on

automated methods to rank resources and their component data items as well as further refinement of the methods for toponym equivalence determination.

Acknowledgements

This research was supported by funding from the European Commission project TRIPOD (IST-FP6-045335) and from the UK Ordnance Survey.

References

1. Goodchild, M.F. and L.L. Hill, Introduction to Digital Gazetteer Research. International Journal of Geographic Information Science, 2008. 22(10): p. 1039-1044.
2. Goldberg, D.W., J.P. Wilson, and C.A. Knoblock, Extracting Geographic Features from the Internet to Automatically Build Detailed Regional Gazetteers. International Journal of Geographical Information Science, 2009. 23(1): p. 93-128.
3. Hastings, J.T., Automated Conflation of Digital Gazetteer Data. International Journal of Geographical Information Science, 2008. 22(10): p. 1109-1127.
4. Keßler, C., K. Janowicz, and M. Bishr. An Agenda for the Next Generation Gazetteer: Geographic Information Contribution and Retrieval. in 17th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems, ACM-GIS 2009. Seattle, Washington, USA.
5. Mikheev, A., M. Moens, and C. Grover, Named Entity Recognition without Gazetteers, in In Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics. 1999. p. 1-8.
6. Uryupina, O., Semi-Supervised Learning of Geographical Gazetteers from the Internet, in HLT-NAACL 2003, Workshop on Analysis of Geographic References, Alberta, Canada. 2003. p. 18-25.
7. Christen, P. and T. Churches. A Probabilistic Deduplication, Record Linkage and Geocoding System. in ARC Health Data Mining Workshop, Canberra, AU, The Australian National University 2005.
8. Hollenstein, L., Capturing Vernacular Geography from Georeferenced Tags, Msc Thesis, in Department of Geography, University of Zurich. 2008.
9. Rattenbury, T. and M. Naaman, Methods for Extracting Place Semantics from Flickr Tags. ACM Trans. Web,, 2009. 3(1): p. 1-30.
10. Popescu, A., G. Grefenstette, and P.-A. Moëllic, Gazetiki: Automatic Creation of a Geographical Gazetteer, in JCDL '08: Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries. 2008, ACM: New York, NY, USA. p. 85-93.
11. Wiederhold, G., Mediators in the Architecture of Future Information Systems. Computer IEEE Computer Society Press, 1992. 25(3): p. 38-49.
12. Callan, J., Distributed Information Retrieval. In: Advances in Information Retrieval. Kluwer Academic Publishers, p: 127-150. 2000.
13. Smith, J.M., P.A. Bernstein, U. Dayal, N. Goodman, T. Landers, K.W.T. Lin, and E. Wong. Multibase - Integrating Heterogeneous Distributed Database Systems. in AFIPS National Computer Conference. 1981.
14. Gupta, A., R. Marciano, I. Zaslavsky, and C.K. Baru, Integrating GIS and Imagery through XML-Based Information Mediation, in International Workshop on Integrated Spatial

Databases, Digital Images and GIS, P.Agouris and.A. Stefanidis, (Eds) 1999, Springer LNCS 1737, p. 211-234.
15. Zaslavsky I., Gupta A., Marciano R., and B. C., Xml-Based Spatial Data Mediation Infrastructure for Global Interoperability (Available from http://www.gsdi.org/capetown/program.htm). in 4th Global Spatial Data Infrastructure Conference. 2000.
16. Gupta, A., Memon, A., Tran, J., Bharadwaja, R. P., and Zaslavsky, I. , Information Mediation across Heterogeneous Government Spatial Data Sources, in Annual National Conference on Digital Government Research. 2002, Digital Government Society of North America: Los Angeles, California. p. 1-6.
17. Diggle, P.J., J. Besag, and T.J. Gleaves, Statistical Analysis of Spatial Point Patterns by Means of Distance Methods. Biometrics, 1976. 32(3): p. 659-667.
18. Levenshtein, V.I., Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. Soviet Physics Doklady, 1966. 10(8): p. 707-710.
19. Russell, R. and M. Odell. The Soundex Indexing System, National Archives and Records Administration. 1918 Available from: http://www.nara.gov/genealogy/coding.html.
20. Sinnott, R.W., Virtues of the Haversine. Sky and Telescope, 1984. 68(2): p. 159-162.