An Agenda for the Next Generation Gazetteer: Geographic Information Contribution and Retrieval

Carsten Keßler
Institute for Geoinformatics
University of Münster
Münster, Germany
c.kessler@uni-muenster.de

Krzysztof Janowicz
Institute for Geoinformatics
University of Münster
Münster, Germany
janowicz@uni-muenster.de

Mohamed Bishr
Institute for Geoinformatics
University of Münster
Münster, Germany
m.bishr@uni-muenster.de

ABSTRACT

Gazetteers are key components of georeferenced information systems, including applications such as Web-based mapping services. Existing gazetteers lack the capabilities to fully integrate user-contributed and vernacular geographic information, as well as to support complex queries. To address these issues, a next generation gazetteer should leverage formal semantics, harvesting of implicit geographic information – such as geotagged photos – as well as models of trust for contributors. In this paper, we discuss these requirements in detail. We elucidate how existing standards can be integrated to realize a gazetteer infrastructure allowing for bottom-up contribution as well as information exchange between different gazetteers. We show how to ensure the quality of user-contributed information and demonstrate how to improve querying and navigation using semantics-based information retrieval.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Spatial databases and GIS; H.3.3 [**Information Search and Retrieval**]: Retrieval models; H.5.3 [**Group and Organization Interfaces**]: Web-based interaction

General Terms

Gazetteer, Semantics, Trust, Volunteered Geographic Information

1. INTRODUCTION AND MOTIVATION

Digital gazetteers are directories containing triples of place names (N), geographic footprints (F), and feature types (T) for named geographic places [26]. In general, they offer at least two functions, one which maps from place names to footprints ($N \rightarrow F$) and one from place names to feature types ($N \rightarrow T$). The feature types are mostly organized in semi-formal thesauri with natural language descriptions. In the context of gazetteers, a clear distinction is made between place as a social construct based on perceivable characteristics or convention [11], and the actual real-world feature it refers to [21]. The Open Geospatial Consortium (OGC) views gazetteers as special profiles[1] of Web Feature Services (WFS) [45]. According to this definition, gazetteers are more than just yellow pages for places, they are the subset of web feature services that deal with named places. Gazetteers are a fundamental building block for applications such as web-based mapping services, spatial search engines, and geoparsers. Working under the hood of these applications, they enable queries such as *rivers in Washington*. While most existing gazetteers are able to respond to such comparably simple queries, we argue that they bear potential for improvement in a number of ways. For instance, they cannot respond to complex queries spanning across different gazetteers and typing schemas. Additionally, they lack the functionality to handle evolving data of differing quality, e.g., user generated content. In this case, the focus shifts to vernacular names as well as dynamic and small-scale feature types such as pubs.

To solve these challenges, a next generation gazetteer has to support the following functionality:

- **Harvesting and integration.** While most gazetteers are maintained by single authorities, the next generation of gazetteers should also incorporate local, small-scale features and feature types maintained by the community. This requires a distributed approach for gazetteer infrastructures. Moreover, novel harvesting and extraction strategies are required to derive information about places from implicit and explicit volunteered geographic information and integrate it based on shared typing schemas.
- **Assessing fitness for purpose.** Gazetteers are used for different purposes. While some applications require stable, high-quality content provided by legal authorities, emerging (e.g., mobile) services require timely and user-centric information. Consequently, the intended application directly influences which data fits its purpose. This problem is especially apparent for user generated content, where the contributors, accuracy, and lineage differ frequently across data sets. Recent research indicates that trust rankings can be employed as proxies for the fitness for purpose.

[1]The specification of the WFS gazetteer profile is an ongoing task, see http://www.opengeospatial.org/projects/groups/wfsgaz1.0swg.

- **Retrieval, querying, and navigation.** The feature type lists or semi-formal feature type thesauri employed in most gazetteers hamper discovery, navigation, and complex queries since they do not support logical inference (e.g., satisfiability checking, subsumption and similarity reasoning) but consist of ambiguous plain text descriptions. Moreover, the chosen type hierarchy and *related-to* relations are static and not always intuitive. Switching to feature type ontologies would make it possible to introduce arbitrary relations, uncover implicit relations (such as similarities between types), ensure consistency, and support complex queries as well as a more intuitive navigation between feature types.

In this paper, we discuss the recent challenges in gazetteer research, review existing approaches, and point out the missing pieces required for a next generation gazetteer infrastructure (NGGI). We focus on a distributed approach for the integration of volunteered geographic information as well as semantics-based retrieval and navigation. Section 2 deals with the question of how to harvest and integrate explicit and implicit user-generated geographic information from the web. Section 3 elaborates on the role of trust in this process, followed by an analysis of the requirements for and benefits of formal semantics for gazetteers in Section 4. We conclude the paper with directions for future research.

2. INCORPORATION OF VGI

The amount of volunteered geographic information available on the social web has been soaring since the advent of geotagging, mash-ups and Google Earth. Compared to other projects building on user contributed content, such as Wikipedia, volunteered geographic information (VGI) [5, 20] is not yet widely accepted as a valuable information source. VGI often contains knowledge about vernacular places and their names [49]. In many cases, such information is produced by locals who can provide timely updates. The Open Street Map project[2] is an impressive showcase of the potential of VGI which outdoes commercial mapping sites in terms of update frequency and thematic scope (e.g., bike paths or hiking routes) in many regions. Other sources for partially high-quality VGI are GeoRSS feeds, KML or GML files, and archives such as Microsoft's Bing maps collections and the Google Earth community. The underestimation in usefulness of VGI becomes more evident when the notion of VGI is extended to information that contains geospatial information as an add-on, such as geotagged photos.

Existing gazetteers such as the Alexandria Digital Library (ADL) gazetteer [27] or GeoNames[3] have built their own collection of place names, spatial references, and typing schemas, which were either created from scratch or inherited from government agencies. Building and maintaining such a sizable collection can be tedious, especially if any suggested updates have to be verified and completed manually. To overcome this problem, we suggest a semi-automatic approach to include VGI in gazetteers. Including such information does not only help to keep the gazetteer up to date, it also extends it with vernacular names that are not officially used.

[2] http://www.openstreetmap.org/
[3] http://www.geonames.org/

A next generation gazetteer should leverage volunteered geographic information as an additional information source to improve the completeness and update frequencies; the challenge in doing this is to (1) find robust mechanisms to harvest VGI that filter out the inherent noise in such information created on the social web, and (2) to align the extracted geographic information to existing gazetteers.

2.1 Harvesting Implicit Geographic Information

A number of approaches have been developed in different research fields towards the capturing and representation of named places. In [43], bottom-up geographic information was collected from participants who were asked to outline the area they considered *downtown Santa Barbara* on a paper map. Although it was not possible to derive a crisp border from the participants' sketches, there was still a solid agreement on the core downtown area. A linguistic approach is introduced in [37], using documents harvested via Web search. It is based on co-occurrence of place names and produced comparable results to those of the previously mentioned method using paper maps. The opportunity to use large document bodies from the Web allows for more rapid analyses on a broader basis. A similar approach is presented in [25] to enable querying and mapping of non-spatial terms (e.g., *jobs in Seattle*). Likewise, the notion of *geographic semantic relatedness* [24] allows for mapping non-spatial terms to related places based on Wikipedia entries. In [44], a novel method is introduced for the integration of remote sensing imagery into gazetteers to automate geographic data management. In another linguistics-based approach presented in [50], a bootstrapping algorithm is applied to automatically classify places into predefined categories (e.g., *city*, *mountain*). Although the machine learning techniques employed in this research were provided with only 100 examples per category, they still yielded a high precision of about 85%.

While these approaches are all based on place names in texts, implicit VGI may also be hidden in content such as geotagged photos or blog posts. Since such content already provides coordinate pairs in combination with keywords, their positions and assigned tags can be analyzed with different approaches such as Delaunay triangulations and Voronoi diagrams, alpha shapes [15] or egg and yolk representations [12, 51]. As such, a plethora of methods has already been developed for harvesting implicit and explicit geographic information from the web. Some of them are already used in practice, such as for Yahoo!'s collection of shape files[4] extracted from the large number of geotagged photos (over 100 million at the time of writing of this paper) uploaded to their Flickr community.

These purely geometrical approaches produce useful results for areal features such as cities. However, they cannot be applied stand alone for the extraction of place names for gazetteers. To extract full gazetteer entries, place names, types and geographic footprints are needed. Both place names and eventually types can be found in the resources' tags, which are a potential source for a folksonomy [46]. Extraction algorithms cannot distinguish between tags which are place names and other tags (e.g., *party*, *Canon*). Place

[4] http://code.flickr.com/blog/tag/shapefiles/

names can only be identified by analyzing the spatiotemporal distribution of the tags[5]: A tag referring to a long-term place name forms a spatial cluster (or multiple clusters for different places going under the same name), but it does not cluster temporally. While the spatial clustering ensures that non-spatial terms are filtered, the equal temporal distribution filters out short term (and potentially recurring) tags for events and other dynamic places. The temporal dimension hence plays a crucial role for the place name extraction process. Gruber [22] suggests to model a tag as a relation between a place name N, a unique identifier pointing to the annotated resource R, the author U and the creation date (and time) D. This definition needs to be extended by the location coordinates L to match it our notion of a geotagged resource as required for our extraction approach. On the social dimension, it must be assured that a place name is used across the community: if a tag fulfills the spatiotemporal requirements, but it is only used by a marginal number of people for the resources in a spatial cluster, it should not be considered a candidate place name. Any approach working on the tags can be improved by standard preprocessing methods such as porter stemming and filtering of stop words.

A predefined typing scheme (see Section 4) is indispensable for identification of place types in the tags, since no spatiotemporal clustering behavior can be defined for them. Accordingly, bottom-up generation of the type hierarchy and definitions is clearly not feasible at this point. It is, however, possible to define required topological relations for the different feature types: an island has to be *inside* water, a train station has to be *next to* a railway line, etc. The set of allowed relations may also contain legal restrictions – in the US, coffee bars (and other private businesses) cannot be on state-owned land such as in public parks, for example. The specification of such relations would help to solve the problem that the geometric approaches are generally "blind" for wrong tags, which occur frequently when users tag a photo with the place shown in the picture, but it does not coincide with the place where the picture was taken (see Figure 1). Moreover, users often tag all pictures in an album with a place name (e.g., *Seattle*), even if some of them have been taken on the way to – and not *in* – Seattle, for example. Finally, on a very small scale, erroneous GPS positions can hamper the extraction of features. Enforcing rules for allowed topological relations is a promising approach towards filtering out geotagged resources that violate these rules.

2.2 Handling and Aligning VGI

Aligning volunteered geographic information to an existing gazetteer is straightforward if a piece of VGI commits to the same feature type definitions as the gazetteer (or gazetteer infrastructure, see Section 4.1). New strategies are required, though, to check the validity of user-generated entries over time. VGI shows its strengths in dynamic situations, e.g., in case of wild fires. In such cases, people depend on timely updates to protect themselves and their property. Official agencies producing geographic information are often unable to cope with these requirements, so that people have turned to maps generated by the community to inform themselves about the current situation.

[5]Given that no existing place name directories are used during the extraction.

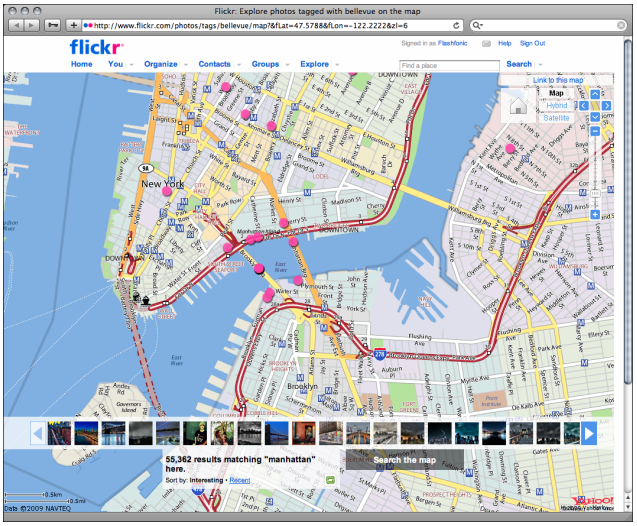

Figure 1: The pink dots show the locations of photos tagged with *Manhattan*. Some have been taken on the Queens side of East River and are thus tagged wrong with respect for the given task. Screenshot from http://www.flickr.com/photos/tags/manhattan/map?&fLat=40.7044&fLon=-73.973&zl=4.

Integration of such information into gazetteers hence requires a versioning mechanism that enables the annotation of entries with time spans for which they were (or still are) valid. In fact, a mechanism would be required to capture the *continuous* change of places' footprints over time; recent research investigates whether this would require places to be modeled as perdurants [31]. In Section 3, we discuss how the temporal dimension causes trust in information to decay over time. Extraction of the dates for which short-term places are valid is straightforward, using creation (e.g., in a photo's Exif data) and publication dates. Table 1 shows the corresponding extended gazetteer entries. Storing temporal information, such as in the Getty Thesaurus of Geographic Names, turns gazetteers into a valuable resource for research on historic places [30]. Moreover, short-term places like wildfires or demonstrations can be added to complement the static information present in gazetteers today with information about places of a more dynamic nature. Their identification also follows the clustering approach outlined in the previous section, except that short-term places (or events) cluster both spatially and temporally.

Table 1: Extended gazetteer entries. The footprints contain the GML code for polygons occupied by the features.

Name	Type	Footprint	Valid
Constantinople	capital	`<gml>`	306-01-01—1453-05-29
Montecito Tea Fire	wildfire	`<gml>`	2008-11-13—2008-11-18
Münster	city	`<gml>`	793-01-01—now

Figure 2 shows a generic overview of the extraction process of bottom-up gazetteer entries from geotagged resources on the web. So far, we have followed the naïve assumption that every geotagged resource on the web is equally useful for this extraction process. However, as any other kind of geographic information, VGI can vary significantly with respect to its quality. While professional GI usually comes (or at least should come) with metadata on accuracy, lineage, etc., such information is hardly provided for geotagged resources on the web. In the following, we discuss how trust can be employed as a proxy for information quality.

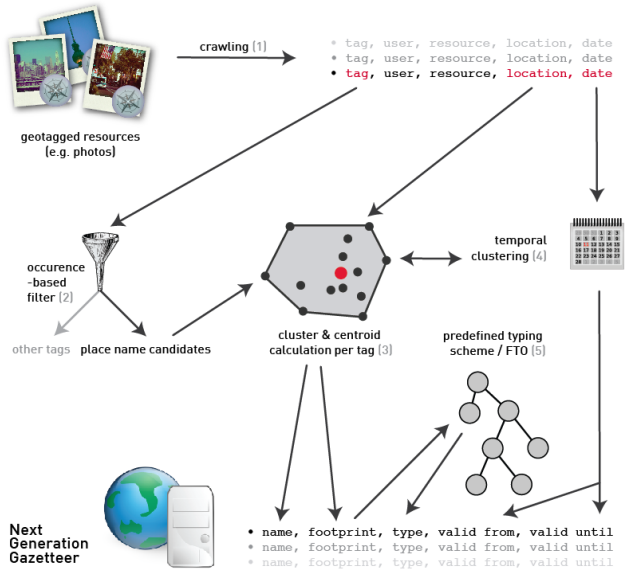

Figure 2: Augmenting gazetteers with VGI starts with crawling geotagged resources from the web (1). The tags are filtered based on occurrences to retrieve toponyms (2). For each place name, region and centroid are calculated based on the coordinates attached to the corresponding resources (3). The check for temporal clusters (4) allows for a distinction between long-term and short-term places. To complete the gazetteer entries, every place name is categorized using a predefined typing scheme (5).

3. A MATTER OF TRUST

Trust is a widely studied phenomenon. From computer science, sociology, social psychology to economics, and many other fields trust has been a central concept [1, 48, 42, 9]. When discussing the issue of trust in gazetteers we can distinguish between two main problems, which entail different types of trust:

- **Trust for distributed gazetteers.** In distributed systems like the internet, the communication pattern involves several entities which did not interact with each other previously. As such, a trust model is required to ensure that otherwise unknown entities can safely interact together and to exclude non-benign entities from such interaction. Trust here refers to security, authentication, authorization and such notions as often discussed in distributed computer literature (e.g., [1, 8]). A distributed gazetteer infrastructure would need similar models to those employed in distributed computing to ensure that the system cannot be maliciously manipulated, accessed or undermined. Such models are not in scope for this paper, however, further research is required to select a proper distributed trust model for a next generation gazetteer infrastructure.
- **Trust for VGI in gazetteers.** In this paper we have a closer look at the problem of trusting VGI. In a gazetteer where user contributions are to be taken as a data source to enrich the gazetteer, the problem of information quality is pressing. Contributors are not equally experienced about the places they report information about, neither are they experienced in a given feature type schema of a certain gazetteer. The veracity of the information they provide needs to be verified before this information can be deemed trustworthy for inclusion into a gazetteer and consequently being served to potential users.

In the next section we look into the problem of VGI veracity and discuss trust as means to cope with the lack of traditional quality criteria in VGI.

3.1 Trusting VGI in Gazetteers

Web 2.0 unleashed the potential of crowd sourcing of information. Users were given the tools to contribute their knowledge to public information repositories that expanded exponentially, and the amount of knowledge collected posed challenges for information quality. Given the large flow of information from user contributions, filtering through this information to extract useful information entities while discarding fraudulent or less credible information becomes a central issue. In [5] the authors address the specific problem of the quality of the potentially enormous volume of geospatial information on the social web. Such phenomenon of user contribution was termed collaboratively contributed geospatial information (later, the term VGI was coined in [20] to describe the same phenomenon). The concern was the emergence of large scale mapping applications where local knowledge about places was being collected by users and contributed to these applications without any guarantees to its veracity. VGI quality was clearly a main concern hampering adoption of VGI in scientific or commercial applications. In user contribution-based applications the traditional quality criteria (i.e., accuracy, completeness, consistency and lineage) are generally missing. Novel ways would need to be developed to assess the quality of VGI, one that utilizes the existing information and that requires minimal additional information to be provided by the VGI contributors.

In [5] the authors proposed to use trust and reputation as a proxy for VGI quality. Such vision is based on the premise that trusted users tend to provide more useful information compared to less trusted users (see e.g., [18]). As such, we say if a trust-rated GI entity is useful and relevant to a large group of users, it can be said that it is of acceptable quality in a more objective sense. Thus, by determining trust values for VGI entities, we can assert some level of satisfaction about their inherent quality. Note that quality refers to information quality (as opposed to data quality) which is more concerned with fitness for purpose [10].

To be able to discuss the matter of trust in this context, one would have to define trust. The problem here is the plethora of different definitions of trust, probably due to the fact that there are many types of trust [17]. The definition we adopt here is that trust is a bet about the future contingent actions of others [48]. That is to say that when users provide VGI at the present moment that is proven to be of good quality, then we can safely assume that future contributions from the same user will be of similar quality. A question of course arises as to whether the VGI provided has to be about the same area or in the vicinity of the previous VGI, since it is evident that the quality of people's VGI contributions will differ with their level of expertise about the nature of the information as well as their proximity from the location they are reporting from (see e.g., [6, 4]). Additionally, the temporal nature of VGI implies that the usefulness of some information types will decay over time. Thus, a trust model for VGI has to be spatially and temporally sensitive [5].

3.2 VGI Trust Models for Gazetteers

Wikipedia shows that as the prominence of a certain user contribution environment increases, the more it becomes subject to malicious behavior[6], and gazetteers will be no exception. In our view the aim of trust models for VGI in gazetteers is to filter through the large flow of information coming from the contributors of information to the gazetteers. We need to be able to scan the enormous amount of user contributions and extract potentially useful information (i.e., trusted information) while discarding incorrect, inaccurate and fraudulent information. However, little work has been done to address the problem of trust in VGI. In [6, 4, 5] the authors discuss the problem of trust in VGI and introduce models that are spatially sensitive to accommodate the specific needs of VGI. In [16] the authors discuss theoretical foundations of information credibility as a basis for discussing credibility of VGI.

Based on our vision of next generation gazetteers, we introduce a set of requirements regarding the VGI trust models that need to be embedded within the gazetteers infrastructure. The idea of such trust models centers around what we call Crowd Sourcing the Filter (CSF)[7]. The idea is to allow the users to assert the fitness for purpose for the VGI they are using from the gazetteers, expressed as trust ratings on both the content and contributors. In the background, spatio-temporally sensitive trust models will compute overall trust values for information entities within the gazetteer and help users find trusted VGI while warning them against fraudulent, malicious, or incorrect information. The set of basic requirements are listed below:

- **Minimal user effort.** A trust model for VGI in gazetteers should not increase the entry barrier to users of the system. Web 2.0 approaches require ease of use in order to encourage user contributions. The trust model must depend on minimal metadata requirements concerning the users and their contributions without requiring elaborate metadata collection.
- **Intuitive trust information capturing.** Gazetteer information consumers must be able to easily assert their level of trust in the information they have already used. The user feedback system employed in the gazetteer should make it easy for users to rate the information they have used on clearly defined scales. In [4] an intuitive 0-10 scale was introduced, and is grounded in theoretical work on quantifying trust [17]. The trust ratings provided are then used by models as discussed earlier to compute overall trust values for specific information entities.
- **User reputations integration.** The trust models should integrate computational modules of user reputations. Reputation here is defined as the collective opinion of the system users about the competence of a specific user in providing quality VGI. Currently the third author is developing a model that allows user reputations to develop slowly with each successful contribution they make, while their reputations is tarnished more rapidly if mistakes are committed. Such approach resembles our everyday experience with people's reputation. Reputation takes time and effort to build on part of the individual, but is easily tarnished if abused.
- **Trust model transparency.** To increase the overall trust in the system by the users, the model has to be transparent to the users. By this we mean that users should be able to tell how the model ratings of information are generated and maintained, which enhances perceptions of the context in which the ratings were generated.
- **Experience, and spatio-temporal sensitivity.** The model should account for user expertise as a defining factor for how much a user should be trusted. For example, some users will be experienced about Berlin at a certain time. Their experience is spatial in nature, but it is also strongly temporal since experience about places decays in the long run if the person is not regularly experiencing the place. As such we assert that spatiotemporal trust models [5] are essential for the success of VGI filtering for gazetteers.
- **Collecting and managing provenance.** Due to the nature of VGI systems, the basic elements of provenance such as purpose of collecting the information or the original author's level of expertise are difficult to collect. Trust models for gazetteers will thus need to capture implicit provenance information from the users. Such information includes, but is not limited to, a user's interaction history with the system, trust ratings, cumulative user reputations, time stamps of contributions or modifications and user profile information. These provenance elements will need to be integrated within trust models as weighting factors when performing trust computations for VGI entities.

Both models of trust as well as the VGI extraction rely on semantic descriptions of the gazetteers' contents. The following section discusses this semantic enablement.

[6]This example shows a prominent Wikipedia fraud case. This (and other) incidents prompted Wikipedia administration to research potential trust models to secure the future of Wikipedia: http://www.nytimes.com/2007/03/05/technology/05wikipedia.html?_r=1&ref=business.

[7]See e.g., Ushahidi project http://www.ushahidi.com/.

4. SEMANTIC ENABLEMENT FOR GAZETTEERS

The abstraction from features to feature types is a core component of all major web gazetteers. It enables type-based queries such as *list all rivers in the selected map extent* in the first place. Complex queries consisting of relations beyond spatial or administrative containment and especially reasoning on the feature types, however, are currently not supported [29]. While the required information to respond to queries such as for *lakes in wildlife reserves near Seattle* should be contained in most gazetteers, it is currently not possible to pose such queries. This lack of advanced query functionality stems from missing reasoning capabilities. Hence, from a functional point of view, this problem can be solved by introducing formal feature type specifications and relations between them (and their instances). A similar challenge as discussed for gazetteer data also applies to the typing schemas employed in current gazetteers. Most gazetteers have developed their own schemas. Since these are realized as lists or thesauri [32] and are not based on formal feature type definitions, automatic and meaningful mappings between different gazetteers are difficult to achieve, even with manual intervention. A recent approach for such a gazetteer conflation was presented by Hastings [23].

One example for the lack of formal semantics is the feature type *offshore platform* in the ADL feature type thesaurus[8]. Same as reservoirs and other types, offshore platforms are defined as narrower terms [3] of hydrographic structures which in turn are described as man-made bodies of water. Consequently [3, p. 46-48], all offshore platforms are water bodies. A search for *man-made bodies of water near Santa Barbara, CA* using the term *hydrographic structures* in the ADL gazetteer would therefore also list several offshore platforms. Note that we do not blame the creators of such thesauri, but argue that the informal or semi-formal character of feature type lists and thesauri can easily lead to such errors; see [32] for a detailed discussion. A semantic engineering [40] approach, e.g., based on ontologies, would largely reduce such errors. First, it requires a formal, i.e., unambiguous, specification of the feature types, and second logical inference can be used to resolve inconsistencies.

4.1 Feature Type Ontology and Distributed Gazetteer Infrastructure

We suggest developing a feature type ontology (FTO) that provides formal definitions for the basic feature types found in gazetteers. This would allow single gazetteers to commit to this domain ontology, extend it and develop their own – potentially different – notions of the feature types. The conversion process from feature type thesauri to feature type ontologies has been discussed and demonstrated[9] by Janowicz and Keßler [32]. Based on the research agenda [19] set up during the *Digital Gazetteer Research & Practice Workshop*, the main motivation (beside avoiding errors caused by informal type definitions) for semantically enabled gazetteers is to facilitate an interoperable gazetteer infrastructure as shown in Figure 3 [32, 33].

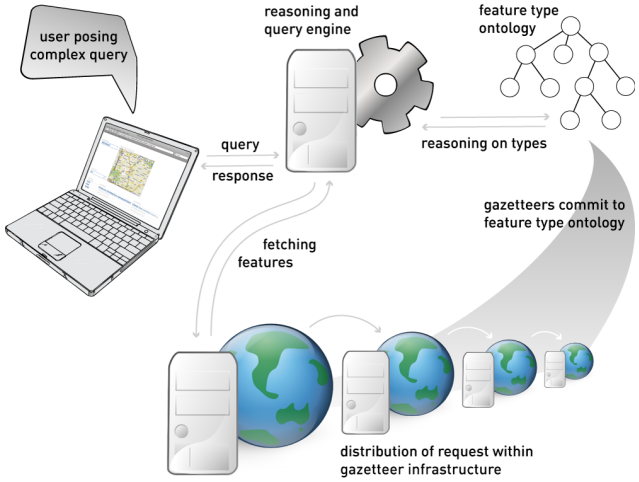

Figure 3: A distributed gazetteer infrastructure.

In such an infrastructure, single gazetteers can be either maintained by local authorities or even volunteers using the VGI-based approach introduced in section 2. Queries can be propagated across different gazetteers, which results in a DNS-like (Domain Name Service) infrastructure for the resolution of place names. If one gazetteer does not contain the appropriate dataset or is not responsible for a specific geographic area, the query is redirected to another gazetteer. This approach is not restricted to dividing and selecting gazetteers by the geographic space they cover, thematic aspects or temporal resolution can also be taken into account. For instance, some gazetteers may only contain stable data validated by some authority, while another gazetteers may store new features contributed by volunteers which have not been evaluated so far. Such an infrastructure relies on the notion of linked data [7, 39] spread across gazetteers instead of single and isolated silo-like gazetteers. For instance, the GeoNames gazetteer already offers a RDF-based link data interface and data from Open Street Map could also be converted to RDF triples easily. Taking temporal aspects into account or allowing users to contribute cultural heritage data would require a linked spatiotemporal data gazetteer [31] infrastructure. From a technical point of view a distributed gazetteer infrastructure can be set up using RDF triple stores, RDF scrapers, and the SPARQL Protocol and RDF Query Language (SPARQL); see [2] for details. The definition of the domain-level feature type ontology and the alignment [47] of the local ontologies requires several steps:

First, the domain-level ontology needs to be generic enough to act as reference for all local gazetteer ontologies. It is unlikely that this can be reached on the level of concrete types such as *reservoir*. Instead, the ontology should define the basic vocabulary for specifying types. This direction was also taken by one of the most successful top-level ontologies, the CIDOC Conceptual Reference Model (CRM) used in the domain of cultural heritage [13]. The strength of CIDOC CRM lies in defining activities (as relations and through reification) as core building blocks instead of concrete exhibit types such as *painting*. Following this idea, a domain-level feature

[8] http://www.alexandria.ucsb.edu/gazetteer/FeatureTypes/ver070302/

[9] A direct mapping from the ADL feature type thesaurus to an OWL-lite version, as well as several more expressive demo feature type ontologies specified in OWL-DL can be downloaded from http://sim-dl.sourceforge.net.

type ontology should not define *parking lot*, *train station*, or *harbor*, but introduce the notion of *transportation* and *transfer point*; see [41] also for a detailed discussion. It is up to the local, application specific gazetteers to align their types and use the building blocks introduced on domain level for their specification. Note that gazetteers use feature types to provide extended query and navigation capabilities, not to run simulations, e.g., for flood prediction. Consequently, gazetteers provide a simplified and restricted view on feature types which makes their alignment and mapping easier.

Figure 4: Conceptual user interface design for the recommender system.

Second, integrating volunteered geographic information into gazetteers relies on a shared understanding of the tags used to describe pieces of information. The lack of tools that support non-professional users in describing their content with common domain vocabulary is one of the main reasons for the wasted potential of VGI. Unlike GI professionals, lay users cannot be expected to provide extensive, ISO-compliant metadata [28] for their contributions that are required to make them findable. Instead, we propose to focus the tagging process most users are already familiar with on the vocabulary of the feature type ontology during the submission of their content. The key component to achieve this goal is a recommender system for semantic annotations. Based on the concepts in the FTO, this system supports novice users in providing accurate annotations. Recommendations are generated based on the content's spatiotemporal and thematic characteristics, as well as the user's history of interaction with the system, e.g., feature types she has looked up previously through the retrieval interface in Figure 5. The implementation of such a system at the level of popular places for the collection of VGI[10] would greatly facilitate the extraction and alignment described in Section 2.

The work-flow for submitting and tagging VGI using the recommender system as shown in Figure 4 consists of 4 steps:

- Uploading the contents or providing a valid URL to online contents (e.g., a geotagged photo or a KML file).
- Automatic analysis of the contents. This step differs depending on the type of content; for photos, the Exif data are analyzed, for KML files, the file contents (keywords, descriptions, spatial extent of layers) can be parsed. This step results in a set of keywords associated with the content.
- Augmentation of the automatically extracted information with similar concepts from the FTO. The context [38] for the similarity search is extracted from the user profile.
- Presentation of the augmented results to the user ranked by relevance. The user can choose which of the recommendations she wants to adapt to annotate the content.

Both annotation and retrieval hence rely on formal type definitions. The following example illustrates the difficulty to define concrete types on the domain-level and motivates the need for a generic FTO to enable interoperability between gazetteers. The ADL feature type thesaurus describes *countries* as *[t]erritory occupied by a large group of people organized under a single, usually independent government, and recognized as a country internationally*, and declares *nations* as non-preferred term, i.e., *countries* should be used when querying for *nations* [3]. In contrast, the Getty Thesaurus of Geographic Names (TGN), prefers *nations* and uses *countries* only for special cases such as the divisions of the United Kingdom (e.g., Scotland, Britain, etc.). As result, a query for countries returns 165 features in ADL while TGN lists 11 countries. The major challenge is not the fact that conceptualizations differ across gazetteers but that the lack of formal specifications does not allow to discover and resolve such differences (semi-)automatically [32].

4.2 Retrieval, Querying, and Navigation

While section 4.1 illustrates the need for semantics to enable interoperability between gazetteers, this section focuses on the front-end, i.e., the user interface and how reasoning can improve the query capabilities.

Most existing user interfaces are bound to a specific gazetteer (protocol). In contrast, the proposed gazetteer infrastructure supports the development of independent interfaces. A single interface can be used to query and navigate through data from different gazetteers. As illustrated in Figure 3, the underlying infrastructure is transparent to the user, i.e., the user does not recognize that the displayed results are derived from multiple sources. While technically this can be achieved by RDF merge via resource URIs[11] and type propagation (see [2] for details), there are two major research challenges related to the development of such interfaces, namely the questions of (1) how to integrate reasoning and query capabilities into a consistent interface that

[10] Such as `http://bbs.keyhole.com`.

[11] This raises some fundamental research questions about the identity of places not discussed here in detail.

Table 2: From classical gazetteers to a next generation gazetteer infrastructure (NGGI).

Dimension	Classical Gazetteers	NGGI
Audience	Human experts	Non-experts and machines
Contributors	Established authorities	User communities
Data integration	Offline harvesting and integration	On-the-fly integration
Trust	Implicit, established authorities	Explicit, community asserted
Provenance	Maintained by authorities	Inferred from data and trust ratings
Types	Large-scale (e.g., city)	Small-scale (e.g., parking lot)
Places	Named (historical) places	Vernacular place names
Inference	Asserted knowledge	Asserted and inferred knowledge
Typing schema	Informal and semi-formal	Semi-formal and formal

hides the underlying complexity from the user; and (2) how to provide additional provenance information to determine whether the data is trustworthy for a given purpose (see Section 3). As data can also be inferred by the reasoner out of existing information, provenance is not restricted to judging the trustworthiness of the contributor; see also [14] for a detailed discussion.

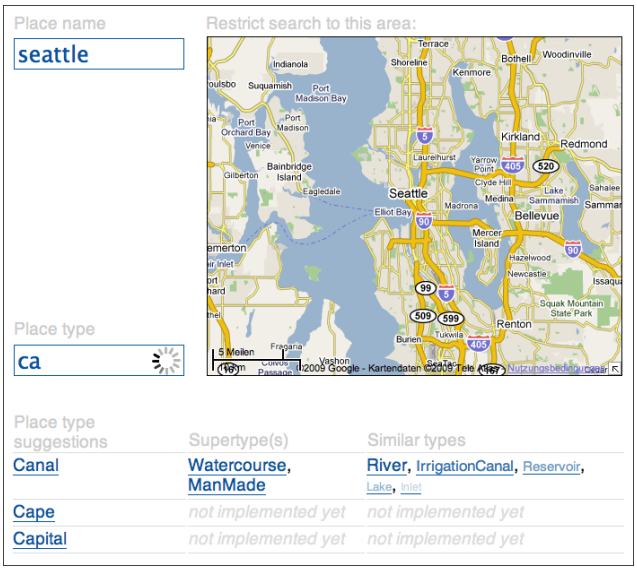

Figure 5: Semantic-enabled gazetteer interface with binding to the ADL gazetteer; adapted from [32].

In previous work, we introduced [33] and evaluated [35] two new gazetteer web interfaces using different semantics-based information retrieval paradigms [36] such as similarity search or query-by-example. For instance, the web interface illustrated in Figure 5 implements a search-while-you-type form for the feature types. It displays type suggestions, the immediate super type (to broaden the search), as well as similar types for horizontal navigation between types[12]. The current interface implements the ADL gazetteer protocol but can be modified to work as interface for the proposed gazetteer infrastructure. How to incorporate provenance and restrict the query by trust ratings is an open issue so far.

[12]Note that the depicted interface is a simplified version used to illustrate the feature type navigation. The implemented version also contains a frame for the retrieved places and is described in [35]. The full source code as well as an online-demo are available at http://sim-dl.sourceforge.net/applications/.

5. CONCLUSIONS AND OUTLOOK

In this paper, we presented a vision for a next generation gazetteer infrastructure. We reviewed existing approaches, identified the challenges and pointed out missing pieces. The proposed infrastructure supports the incorporation of volunteered geographic information, trust-based assessment of information quality, as well as semantics-based retrieval and navigation.

Table 2 shows the transition from classical gazetteers to the proposed gazetteer infrastructure on various dimensions. While most classical gazetteers are used by experts, the focus of the NGGI shifts towards non-expert users and machine-to-machine communication (e.g., for reasoning and harvesting). So far, places have been collected offline from established authorities such as government agencies into silo-like single gazetteers. In the NGGI approach, data can also be extracted from implicit geographic information and stored in distributed gazetteers. Query results spanning over multiple gazetteers are then integrated at query time. In addition, as opposed to classical gazetteers in which trust is implicit in the authorities (i.e., data providers), the NGGI utilizes spatiotemporal trust models as proxies for the quality of VGI (in what we term *crowd sourcing the filter*). Likewise, the data provenance is ensured by authorities; in our approach, data provenance is more challenging due to the nature of VGI. A proposal is made to infer provenance information from history of user interactions such as trust ratings. Including VGI allows us to include small-scale features that were before unattainable in gazetteers, such as shopping districts. While current gazetteers support official and vernacular place names, the NGGI will support the bottom-up evolution of vernacular place names, particularly on the small-scale level. Existing gazetteers are built completely on asserted, i.e., explicitly stated information about the places. Based on formal feature type ontologies, the NGGI can infer additional facts about features and feature types and help overcome the shortcomings of natural language feature type definitions and thesauri used in existing gazetteers.

Building a domain ontology for the NGGI is a challenging task for future work and requires interdisciplinary col-

laboration across different communities. An OGC standards compliant implementation of the infrastructure requires a semantic enablement layer as well as the annotation of gazetteer entries (see [34] for details). While there are a number of existing approaches for harvesting VGI, these need to be made more robust and integrated with models of trust to ensure acceptable levels of information quality. Finally, more extensive user interfaces need to be developed which support the new querying capabilities of the NGGI.

Acknowledgments

This research has been funded by the International Research Training Group on *Semantic Integration of Geospatial Information* (DFG GRK 1498, see irtg-sigi.uni-muenster.de) and the *SimCat II* project (DFG JA1709/2-2, see sim-dl.sf.net), and is part of the semantic enablement initiative at $52^{\circ}North$ (see 52north.org/semantics).

References

[1] A. Abdul-Rahman and S. Hailes. A distributed trust model. In *Proc. of the 1997 workshop on New security paradigms*, pages 48–60. ACM Press New York, 1998.

[2] D. Allemang and J. Hendler. *Semantic web for the working ontologist: modeling in RDF, RDFS and OWL*. Morgan Kaufmann Elsevier, Amsterdam, NL, 2008.

[3] ANSI/NISO. *ANSI/NISO Z39.19 – 2005 Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*. NISO Press, 2005.

[4] M. Bishr. Weaving space into the web of trust: An asymmetric spatial trust model for social networks. In S. Auer, C. Bizer, C. Müller, and A. V. Zhdanova, editors, *Proc. of the 1^{st} Conference on Social Semantic Web*, Leipzig, Germany, 2007.

[5] M. Bishr and W. Kuhn. Geospatial information bottom-up: A matter of trust and semantics. In S. I. Fabrikant and M. Wachowicz, editors, *The European Information Society*, Lecture Notes in Geoinformation and Cartography, pages 365–387. Springer, 2007.

[6] M. Bishr and L. Mantelas. A trust and reputation model for filtering and classifying knowledge about urban growth. *GeoJournal*, 72(3):229–237, 2008.

[7] C. Bizer, T. Heath, and T. Berners-Lee. Linked data – the story so far. *Journal on Semantic Web and Information Systems (IJSWIS); Special Issue on Linked Data*, 2009 forthcoming.

[8] M. Blaze, J. Feigenbaum, J. Ioannidis, and A. Keromytis. The role of trust management in distributed systems security. 1603:185–210, 1999.

[9] V. W. Buskens. *Social networks and trust*. Springer, 2002.

[10] C. Campbell and C. Rozsnyai. Quality assurance and the development of Course Programmes. *Papers on Higher Education. Bucharest, UNESCO, CEPES*, 2002.

[11] R. Casati and A. C. Varzi. *Parts and Places. The Structures of Spatial Representation*. MIT Press, Cambridge and London, 1999.

[12] A. Cohn and N. Gotts. The 'egg-yolk' representation of regions with indeterminate boundaries. In P. Burrough and A. Frank, editors, *Geographical Objects with Undetermined Boundaries*, pages 171–187. Taylor and Francis, London, England, 1996.

[13] N. Crofts, M. Doerr, T. Gill, S. Stead, and M. Stiff. *Definition of the CIDOC Conceptual Reference Model (Manual Version 4.2.1)*, October 2006.

[14] P. da Silva, D. McGuinness, and R. McCool. Knowledge provenance infrastructure. *IEEE Data Engineering Bulletin*, 26(4):26–32, December 2003.

[15] H. Edelsbrunner. Weighted alpha shapes. Technical report, Department of Computer Science, University of Illinois, Urbana, IL, 1992.

[16] A. Flanagin and M. Metzger. The credibility of volunteered geographic information. *GeoJournal*, 72(3):137–148, 2008.

[17] D. Gambetta. *Trust: Making and breaking cooperative relations*. Basil Blackwell New York, 1990.

[18] J. Goldbeck, B. Parsia, and J. Hendler. Trust networks on the semantic web. In *Cooperative Information Agents VII*, volume 2782 of *Lecture Notes in Computer Science*, pages 238–249. Springer, 2003.

[19] M. Goodchild and L. L. Hill. NCGIA summary report: Digital gazetteer research & practice workshop. Technical report, Department of Geography; University of California, Santa Barbara, 2006.

[20] M. F. Goodchild. Citizens as voluntary sensors: Spatial data infrastructure in the world of web 2.0. *Int. Journal of Spatial Data Infrastructures Research*, 2:24–32, 2007.

[21] M. F. Goodchild and L. L. Hill. Introduction to digital gazetteer research. *International Journal of Geographical Information Science*, 22(10):1039–1044, 2008.

[22] T. Gruber. Ontology of folksonomy: A mash-up of apples and oranges. *International Journal on Semantic Web & Information Systems*, 3(2), 2007.

[23] J. T. Hastings. Automated conflation of digital gazetteer data. *International Journal of Geographical Information Science*, 22:1109–1127, October 2008.

[24] B. Hecht and M. Raubal. GeoSR: Geographically explore semantic relations in world knowledge. In L. Bernard, A. Friis-Christensen, and H. Pundt, editors, *The European Information Society*, pages 95–114, Berlin, 2008. Springer.

[25] A. Henrich and V. Lüdecke. Determining geographic representations for arbitrary concepts at query time. In *LOCWEB '08: Proceedings of the first international workshop on Location and the web*, pages 17–24, New York, 2008. ACM.

[26] L. Hill. Core elements of digital gazetteers: Placenames, categories, and footprints. In J. Borbinha and T. Baker, editors, *Research and Advanced Technology for Digital Libraries – 4th European Conference, ECDL 2000 Lisbon, Portugal, September 18–20*, pages 280–290, 2000.

[27] L. L. Hill, J. Frew, and Q. Zheng. Geographic names: The implementation of a gazetteer in a georeferenced digital library. *D-Lib Magazine*, 5(1), 1999.

[28] International Standards Organization. ISO 19115 "Geographic Information – Metadata", 2005.

[29] G. Janée. Rethinking gazetteers and interoperability. In *International Workshop on Digital Gazetteer Research & Practice*, Santa Barbara, CA, December 2006.

[30] K. Janowicz. Towards a similarity-based identity assumption service for historical places. In M. Raubal, H. Miller, A. Frank, and M. Goodchild, editors, *4th International Conference on Geographic Information Science (GIScience 2006)*, volume 4197 of *Lecture Notes in Computer Science*, pages 199–216. Springer, 2006.

[31] K. Janowicz. The role of place for the spatial referencing of heritage data. In *The Cultural Heritage of Historic European Cities and Public Participatory GIS Workshop. 17–18 September 2009*. The University of York, UK, 2009.

[32] K. Janowicz and C. Keßler. The role of ontology in improving gazetteer interaction. *Int. Journal of Geographical Information Science*, 10(22):1129–1157, 2008.

[33] K. Janowicz, C. Keßler, M. Schwarz, M. Wilkes, I. Panov, M. Espeter, and B. Baeumer. Algorithm, implementation and application of the SIM-DL similarity server. In F. T. Fonseca, A. Rodríguez, and S. Levashkin, editors, *Second International Conference on GeoSpatial Semantics (GeoS 2007)*, number 4853 in Lecture Notes in Computer Science, pages 128–145, Mexico City, Mexico, November 2007. Springer.

[34] K. Janowicz, S. Schade, A. Bröring, C. Keßler, C. Stasch, P. Maué, and T. Diekhof. A transparent semantic enablement layer for the geospatial web. In *Terra Cognita 2009 Workshop In conjunction with the 8th International Semantic Web Conference (ISWC 2009)*, (2009; forthcomming).

[35] K. Janowicz, M. Schwarz, and M. Wilkes. Implementation and evaluation of a semantics-based user interface for web gazetteers. In *Workshop on Visual Interfaces to the Social and the Semantic Web (VISSW2009)*, 2009.

[36] K. Janowicz, M. Wilkes, and M. Lutz. Similarity-based information retrieval and its role within spatial data infrastructures. In *5th International Conference on Geographic Information Science (GIScience 2008)*, volume 5266 of *LNCS*, pages 151–167, Park City, Utah, USA, September 2008.

[37] C. B. Jones, R. S. Purves, P. D. Clough, and H. Joho. Modelling vague places with knowledge from the web. *International Journal of Geographical Information Science*, 22(10):1045–1065, 2008.

[38] C. Keßler. Similarity measurement in context. In B. Kokinov, D. C. Richardson, T. R. Roth-Berghofer, and L. Vieu, editors, *6th International and Interdisciplinary Conference, CONTEXT 2007*, volume 4635 of *Lecture Notes in Artificial Intelligence*, pages 277–290, Roskilde, Denmark, 2007. Springer.

[39] D. Kolas and T. Self. Spatially augmented knowledgebase. In *6th International and 2nd Asian Semantic Web Conference (ISWC2007+ASWC2007)*, pages 785–794, November 2007.

[40] W. Kuhn. Semantic engineering. In G. Navratil, editor, *Research Trends in Geographic Information Science*. Springer, forthcoming 2009.

[41] B. Lorenz, H. J. Ohlbach, and L. Yang. Ontology of transportation networks (rewerse-del-2005-a1-d4). Technical report, REWERSE Project IST-2004-506779, 2005.

[42] N. Luhmann. *Trust and Power*. Wiley and Sons, 1979.

[43] D. R. Montello, M. F. Goodchild, J. Gottsegen, and P. Fohl. Where's downtown?: Behavioral methods for determining referents of vague spatial queries. *Spatial Cognition & Computation*, 3(2–3):185–204, 2003.

[44] S. Newsam and Y. Yang. Integrating gazetteers and remote sensed imagery. In *GIS '08: Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, pages 1–10, New York, NY, USA, 2008. ACM.

[45] Open Geospatial Consortium. Web feature service implementation specification 1.1.0. Technical report, 2005.

[46] C. Schlieder. Modeling collaborative semantics with a geographic recommender. In J.-L. Hainaut et al., editor, *ER 2007 Workshops*, volume 4802 of *Lecture Notes in Computer Science*, pages 338–347, Auckland, New Zealand, November 5–9, 2007, 2007. Springer.

[47] P. Shvaiko and J. Euzenat. Ten challenges for ontology matching. In R. Meersman and Z. Tari, editors, *On the Move to Meaningful Internet Systems: OTM 2008*, volume 5332 of *Lecture Notes in Computer Science*, pages 1164–1182. Springer, 2008.

[48] P. Sztompka. *Trust: A Sociological Theory*. Cambridge University Press, 1999.

[49] F. Twaroch, R. Purves, and C. Jones. Stability of qualitative spatial relations between vernacular regions mined from web data. In *Proceedings of Workshop on Geographic Information on the Internet*, Toulouse, France, April 6th, 2009.

[50] O. Uryupina. Semi-supervised learning of geographical gazetteers from the internet. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references*, pages 18–25, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[51] F. Wilske. Approximation of neighborhood boundaries using collaborative tagging systems. In E. Pebesma, M. Bishr, and T. Bartoschek, editors, *GI-Days 2008*, volume 32 of *ifgiPrints*, pages 179–187.