Bottom-Up Gazetteers: Learning from the Implicit Semantics of Geotags

Carsten Keßler, Patrick Maué, Jan Torben Heuer and Thomas Bartoschek

Institute for Geoinformatics, University of Münster, Germany
carsten.kessler|patrick.maue|jan.heuer|bartoschek
@uni-muenster.de

Abstract. As directories of named places, gazetteers link the names to geographic footprints and place types. Most existing gazetteers are managed strictly top-down: entries can only be added or changed by the responsible toponymic authority. The covered vocabulary is therefore often limited to an administrative view on places, using only official place names. In this paper, we propose a bottom-up approach for gazetteer building based on geotagged photos harvested from the web. We discuss the building blocks of a geotag and how they relate to each other to formally define the notion of a geotag. Based on this formalization, we introduce an extraction process for gazetteer entries that captures the emergent semantics of collections of geotagged photos and provides a group-cognitive perspective on named places. Using an experimental setup based on clustering and filtering algorithms, we demonstrate how to identify place names and assign adequate geographic footprints. The results for three different place names (*Soho*, *Camino de Santiago* and *Kilimanjaro*), representing different geographic feature types, are evaluated and compared to the results obtained from traditional gazetteers. Finally, we sketch how our approach can be combined with other (for example, linguistic) approaches and discuss how such a bottom-up gazetteer can complement existing gazetteers.

1 Introduction and Motivation

The amount of geotagged user-generated content on the Social Web has been soaring in the last years. Cheaper and smaller GPS chips as well as easy-to-use tools for manual geotagging have led to a sharp increase, particularly in the number of geotagged photos. The sheer amount of geotagged pictures – currently over 100 million on Yahoo's Flickr service alone[1] – makes them a very attractive source for geographic information retrieval [1,2]. As such, geotagged photos can be regarded as an implicit kind of Volunteered Geographic Information (VGI) [3]. Merging professional data sources with such VGI is attractive for a number of reasons, such as rapid updates and enrichment with data typically not contained in professional data sets. Examples include the extraction of footprints [1] and grounding of vague geographic terms [4] such as *downtown Mexico City*

[1] According to http://blog.flickr.net/2009/02/05/.

or mapping of non-geographic terms [5] to determine the regional use of words like *soda* or *pop* [6].

One promising use of VGI – and geotagged photos in particular – is the enrichment of gazetteers with vernacular names and vague places [7]. Gazetteers have been developed as directories of named places with information on geographic footprints and place types to facilitate geographic information organization and retrieval. Most gazetteers follow a strict top-down approach, i.e., the gazetteer data is administered by the organization running the gazetteer. Only this toponymic authority can add places or place types to the gazetteer and correct erroneous entries, which slows down updates and hampers the inclusion of local and often tacit knowledge. Moreover, in most gazetteers information on geographic footprints is limited to a single coordinate pair, representing the centre of a city, administrative district or street. Extraction of footprints from geotagged information on the web is thus a promising way to automatically generate polygonal footprints for these gazetteer entries. Although a number of approaches have been developed for this task [5,8,9,10], they are hardly implemented in existing gazetteers. Apart from the GeoNames gazetteer[2], which complements its database with geotagged information from Wikipedia, strict top-down management of gazetteers is still prevalent.

In this paper, we present an approach to build gazetteers *entirely* from volunteered geographic information. We discuss the challenges posed by automatically establishing the foundations of such a gazetteer based on geotagged photos harvested from the web. The implemented algorithms for retrieving geotags and clustering the corresponding locations to generate footprints are well-established. However, the *emergent semantics* [11] of such a collection of geotagged photos is still largely unspecified. Hence, the main contribution of this paper will be the formal definition of geotags. We explain the relation between the attached label (tag) and the information objects like a photo, its label's author, as well as creation time and coordinates. We discuss the implicit semantics hidden in this relation, and how gazetteer entries can emerge from collections of such geotags using the presented implementation.

Inferred knowledge about places from a source like geotagged photos – usually tagged with subjective keywords – can be seen as a social knowledge building process [12, chapter 9]. Ideally, this process leads to a representation of the *group cognition* [12] and can thus be regarded as a cognitive engineering [13] process which lets traditional GI applications benefit from the *Wisdom of the Crowds* [14]. Gazetteers exposing the collaborative perspective on place differ significantly from traditional gazetteers with administrative focus [15]. It is thus not the aim of this research to replace today's gazetteers, which have already proven useful for countless applications building on geocoding, geoparsing and natural language processing. Instead, we argue for a separation of these different views into separate gazetteers, which can then be accessed through a gazetteer infrastructure as outlined in [7,16].

[2] See http://www.geonames.org.

In order to demonstrate the feasibility of our approach, we have set up an application which retrieved and processed geotags associated to photos published on Flickr, Panoramio and Picasa[3]. While there is also other geotagged content online such as videos, blog posts or Wikipedia entries, we chose to limit this experiment to photos. Photos are inherently related to the real world, since every photo has been taken *somewhere*. Moreover, as mentioned above, there is already a substantial amount of geotagged photos available online. By analyzing the coordinate pairs attached to the pictures, the time they were taken as well as the tags added by their owners, we are able to compute geographic footprints representing specific keywords. The collection of these keywords, derived from all tags of all retrieved photos, is further analyzed to differentiate between toponyms and tags without spatial relation. We test a repository build up this way with queries for *Soho*, *Camino de Santiago* (Way of St. James) and *Kilimanjaro*. We compare the results to those obtained from the same query on GeoNames. This evaluation focuses on the question whether our bottom-up gazetteer can already take on established gazetteers in terms of completeness and accuracy of geographic footprint.

The next section points to relevant related work. Section 3 introduces a formal definition of geotags and establishes the relation between gazetteers and geotags. Section 4 describes the crawling and filtering approach implemented in the prototype. Section 5 analyzes the results obtained for the three exemplary queries, followed by conclusions and an outlook on potential applications and future work in Section 6.

2 Related Work

This section points to related work from gazetteer research, tagging and bottom-up generation of geographic information.

2.1 Gazetteer Building and Learning

Gazetteers are knowledge organization systems that consist of triples (N, F, T), where N corresponds to the place name, F to the geographic footprint and T to the place type [17]. Since neither N, F nor T are unique, all three components are required to fully represent and unambiguously identify a named place [17, p. 92]. In the context of gazetteers, a clear distinction is made between place as a social construct based on perceivable characteristics or convention [18], and the actual real-world feature it refers to [19]. Feature types are mostly organized in semi-formal thesauri with natural language descriptions. Recent research demonstrates how gazetteers could benefit from more rigorous, formal place type definitions [16] and develops methods for gazetteer conflation [20].

Existing gazetteers have generally been developed based on databases provided by administrative authorities, or by merging existing gazetteers [17]. More

[3] See `http://flickr.com/`, `http://panoramio.com/` and `http://picasaweb.com/`.

recently, the ever-growing amount of information available on the web has been identified as a promising resource of knowledge about named places. Jones et al. [1] introduce a linguistic approach to enrich gazetteers with knowledge about vague places. They use documents harvested via web search and analyze them for cooccurrences of vague place names with more precise co-located places. In another linguistics-based approach presented by Uryupina [21], a bootstrapping algorithm is applied to automatically classify places into predefined categories (e.g. *city*, *mountain*). The machine learning techniques employed in this research enabled a high precision of about 85%, albeit the comparably small training data sets of only 100 samples per category. Henrich and Lüdecke [5] introduce a process based on the results retrieved from a web search engine to derive geographic representations for both geographic and non-geographic terms at query time. Goldberg et al. [22] developed an agent-based system that crawls structured online source such as the USPS zip code database and online phone books. The authors demonstrate that this approach is capable of creating detailed regional, land-parcel level gazetteers with a high degree of completeness.

2.2 User-generated Geographic Information

Online mapping tools with open APIs such as *Google Maps* have enabled the creation of the huge amounts of user-generated geographic information – also dubbed collaborative [23] or volunteered GI (VGI) [3] – in the first place. While this mainly refers to projects like OpenStreetMap[4], we argue that geotags, and more importantly the geographic footprints derived from them, can also be filed into this category. Similar approaches have already been sketched in previous research to derive landscape regions [24] or imprecise definitions of boundaries of urban neighborhoods [8] from such geotagged content. We build on this previous work and show how geographic information collected this way can be processed for the integration with existing gazetteers.

3 What is a Geotag?

We have introduced geotags as particular examples of volunteered geographic information. Before discussing the idea of inferring semantics from the *geotag*, we are going to formally define it.

3.1 Tagging Geographic Information Objects

Humans adding items like pictures to their collections use individual ordering schemes (besides time) to group similar items, keep different items apart and consequently simplify recovery. We order books in our (real) book shelf according to various criteria, including topic, age, thickness, or even color. Such individual preferences re-appear in virtual collections. Using tags – words or combinations

[4] See http://www.openstreetmap.org/.

of words people associate with virtual items – is a well accepted approach to sort items on the virtual shelf. Tags, however, can vary significantly from person to person. The formal definition of a tag therefore has to include both the user and the tagged information object. Gruber [25] suggests to model the tag as the process $Tagging = (L, U, I, S)$, which establishes an immediate relation between the the Label L coming from the User's (U) vocabulary associated to an information Item I. This definition includes a Source S, which enables sharing across applications. In the following, we leave this source aside, since it has no direct impact on the presented approach. The following rule states that, if a label is associated with an item by some user, it is regarded as tag. More importantly, it also states that a tag is always bound to its author and the item:

$$\begin{aligned} \forall l(Label(l) \wedge\ & \exists i(Item(i) \wedge associatedTo(l,i)) \\ & \wedge \exists u(User(u) \wedge createdBy(l,u)) \rightarrow Tag(l)) \end{aligned} \tag{1}$$

Any information object which is inherently hard to classify – basically all non-textual information – requires a solution for its categorization. Tagging is commonly accepted for such contents, such as photos or videos, but also for bookmarks, scientific articles, and many more. In the remainder of this paper, we focus on photos with an identifiable geographical context, e.g. a picture of *La Catedral* in Mexico City. The items in question are therefore related to objects in the geographic landscape [26]. Goodchild's "geographic reality" [27] as formal definition of geographical information takes the spatio-temporal nature of the physical (field-based) reality into account. Humans, however, do not perceive reality as continuous fields. They identify individual objects, either directly or indirectly by looking at photos created by camera sensors.

In this *World of Individual Objects* [26] we only consider particulars (entities existing in space and time) with an observable spatial and temporal extension. Objects on the photo have per se no meaning; in Frank's *World of Socially Constructed Reality* we eventually associate semantics to be able to reference the particulars [28] in spoken language. Such reference can either be a proper name, which is used as unique identifier [29], e.g., *Catedral Metropolitana de la Ciudad de México*, or it links to a category[5] which groups objects sharing common properties, e.g. *cathedral*. We finally identify individual particulars according to their spatial or temporal characteristics, by either referring to complex objects (e.g., *downtown*) or to the homogenous spatial or temporal region the object is proper part of, e.g., *Mexico City*. So far, this follows the definition of gazetteer entries from Section 2.1. The place type T and place name N in the discussed triple (N, F, T) both refer to the particular's semantics, the geographic footprint F on the other hand is related to its spatial extension in physical reality.

The same applies to the labels used to tag a photo, which function as references to particulars in geographic space. The nature of this reference, however, cannot be explicitly described: although it appears to be obvious for the

[5] The reference is then again the proper name of the object's type.

mentioned proper names or category names, most tags associated to photos do not have an objective relation to the geographic object. The label `vacation09` makes perfect sense for the user, who might have sorted all pictures of his Mexico trip using this tag. Once the items are shared, however, such personal tags loose any usefulness. Other examples which have no immediate relation to the depicted particular are labels naming properties of the item itself (e.g. `blue`, `high-resolution`), the process of creating the item (e.g. `nikon`), its potential use (e.g. `wallpaper`), or simply the author's opinion (`interesting`). Note that we assume that it is the user's intention to improve the item's findability; hence, we do not expect to encounter deliberate errors (which is obviously not true in real world settings; we propose an effective solution for this problem in Section 3.3). Once we have identified the references, we can use them to locate the referred-to object in space and time. The following rule makes this dependency between the tag and its role as reference to the depicted particular explicit:

$$\forall l \exists i (\mathit{Tag}(l) \wedge \mathit{Item}(i) \wedge \mathit{associatedTo}(l,i) \wedge \exists p(\mathit{Particular}(p) \wedge \mathit{represents}(i,p)) \rightarrow \mathit{refersTo}(l,p)) \tag{2}$$

The rule does not (and cannot) further specify the reference type. Taking our example of the cathedral, the label `Catedral Metropolitana` is immediately referencing – here as proper name – the particular. We can then further specify the tag as a proper name:

$$\forall l \exists p (\mathit{Tag}(l) \wedge \mathit{Particular}(p) \wedge \mathit{names}(l,p) \rightarrow \mathit{ProperName}(l)) \tag{3}$$

The open question here is obviously how to infer if the label is a proper name and, even more important, how to ensure that it is really the proper name of the depicted geographic object. The clustering and filtering approach introduced in the next sections provides answers to both questions.

Labels like `Mexico` or `Summer 2009` are indirect references. They point to a region containing the particular (spatially and temporally, respectively). The following rule formalizes our assumption, that, if the tag is a toponym referring to a certain geographic region, we can infer that our depicted object is spatially related to that region:

$$\forall l \exists p (\mathit{Tag}(l) \wedge \mathit{Particular}(p) \wedge \mathit{refersTo}(l,p) \wedge \exists r(\mathit{GeographicRegion}(r) \wedge \mathit{names}(l,r)) \rightarrow \mathit{spatiallyRelated}(p,r)) \tag{4}$$

We can only assume that there is a spatial relation between the depicted particular and the place name. By looking only at the labels we cannot infer what kind of spatial (or temporal, for that matter) relation exists, and hence what spatial character this specific label has. In the following section we introduce the concept of a geotag as an extension of the traditional tag. Geotags give us the opportunity to make use of geographic coordinates and points in time to identify the spatio-temporal character of the associated labels.

3.2 A Formal Definition of Geotag

The tagging process establishes the relation between the user, the information item, and the label. If the information item represents one or more geographic objects, the associated label may (but does not have to) refer to either dimension of the depicted object: either its semantics (including a proper name of the individual or category) or its spatio-temporal extension (naming, for example, the containing region). A geotag extends the notion of the tag by adding an explicit location in space and time to the information item. In the case of digital photos, a time stamp with the creation date is usually added by the camera automatically. Geographic coordinates are either provided by built-in GPS modules, or added manually by the user. Building on Gruber's definition of tagging as a relation, we add the time stamp T and the coordinates C to the relation (and omit the source S): $Geotagging = (L, U, C, I, T)$. By extending our rule-based definition of a tag (Eq. 1), the following rule reclassifies a label as a geotag

$$\begin{aligned} \forall l \exists i (& Label(l) \wedge Item(i) \wedge associatedTo(l, i) \\ & \wedge \exists c (Coordinate(c) \wedge associatedTo(c, i)) \\ & \wedge \exists t (Timestamp(t) \wedge associatedTo(t, i)) \\ & \wedge \exists u (User(u) \wedge createdBy(l, u)) \rightarrow Geotag(l)) \end{aligned} \tag{5}$$

Note that we do not assume that a label reclassified as geotag is per se a place name. The tag `blue` is not necessarily related to the depicted object, nor does it have a spatial or temporal character. In our understanding, it is still a geotag, since it is the label used by one user in some occasion to tag an item with an associated location and date. In the following Section 3.3, we introduce an approach which reliably computes whether a label is spatially related to the particular.

3.3 A Clustering Approach to Categorize Geotags

The definition of geotags introduced in the previous section has substantial implications on the conceptual level. An information item is linked to a coordinate and time stamp, and labelled by one or more individuals. If we want to extract one particular aspect, e.g. the spatial coverage of geotags, we have to consider the other four properties as well.

Using the definition of a geotag as the relation $Geotagging = (L, U, C, I, T)$, we use the *tuple relational calculus*[6] [30] in the remainder to specify the queries used to retrieve different kinds of *clouds*. For example, the query $\{g.C | g \in Geotagging \wedge g[L] = L_i\}$ returns the coordinates of all tuples g where the label (the field L) has the value L_i. We call the result of this query a point cloud of a label. A folksonomy – the aggregation of all tags from all users into one (uncontrolled) vocabulary – is then simply formalized as $\{g.L | g \in Geotagging\}$. The

[6] TCR is a concise declarative query language for the relational model, the presented examples can also be expressed in SQL.

resulting tag cloud can also be reduced to the vocabulary of one particular user U_i with the query $\{g.L|g \in geotags \wedge g[U] = U_i\}$. Her spatio-temporal activity – the user's movement across space and time – is queried using the statement $\{g.C, g.T|g \in Geotagging \wedge g[u] = U_i\}$.

We suggest to make use of the point cloud of one label to compute its spatial footprint. A gazetteer build on top of this approach could then return geometries and centroids for proper (potentially unofficial) names of geographical objects. The information we derive from geotags, however, is inherently noisy: many tags do not have an immediate relation to the particular represented by the geotagged item. Only *significant occurrences* of geotags should therefore be considered for this approach. We define one occurrence of a geotag $g = (L_i, U_i, C_i, I_i, T_i)$ as significant if the following two conditions are fulfilled:

1. At least two tuples g_i and g_j exist where $g_i[L] = g_j[L]$, and $g[U_i] \neq g[U_j]$. Since names in geotags are subjective, this rule assures that only names which are used by different persons are taken into account.
2. The spatial distribution $\{g.C|g \in Geotagging \wedge g[L] = L_i\}$ can be clustered.

In the following section we describe the algorithm which applies filters checking for these conditions to extract the relevant candidates for toponyms from the large set of tags. The semantic analysis of the two preceding sections can be easily realized as executable rules, for example expressed in the Semantic Web Rule Language (SWRL) [31]. SWRL supports built-ins, the algorithm presented in the following pages can therefore be integrated as *geotag*:*significant* and used to extend and clarify rule 2:

$$\begin{aligned} &\forall l \exists i (\, Tag(l) \wedge Item(i) \wedge \\ &associatedTo(l, i) \wedge geotag : significant(l) \wedge \\ &\quad \exists p(Particular(p) \wedge represents(i, p)) \rightarrow refersTo(l, p)) \end{aligned} \tag{6}$$

A reasoning engine triggers the execution of the clustering algorithm once it processes the added built-in. The algorithm returns true if the given label is significantly occurring (or false otherwise). Once we have applied the filtering and clustering, our gazetteer can provide the point clouds (and the regions covered by the point clouds) for given place names. For some place names, the clustering process results in multiple clusters (see the example of *Soho* in the following sections). This does not impair the efficacy of the presented approach as long as the clustering algorithm produces reasonable results (which depends mostly on the number of available geotags). For cases such as *Soho*, multiple gazetteer entries are generated.

Although we introduced time as a fundamental component of the geotag, we have not discussed the implications for the targeted gazetteer. With the presented approach, the tag `GEOS 2007` would also be classified as place name. While we cannot discuss this issue here in detail for a lack of space, distinguishing between toponyms and labels naming temporal events can be implemented by applying the clustering approach both to the spatial and temporal dimensions.

3.4 Extraction of Gazetteer Entries

Section 2.1 defines gazetteer entries as triples (N, F, T). This notion has to be further specified for a gazetteer based on geotags. Since, in our case, the underlying data consist of a large collection of photos geo-located with exactly one coordinate pair, the given place name N maps to a point cloud as geographic footprint: $F = \{g.C | g \in \mathit{geotags} \wedge g[L] = L_i\}$. Each point in the cloud represents one significant occurrence of the given place name as tag for a photo. Since the footprint is no longer a single coordinate pair, the gazetteer's mapping from place name to footprint $N \longrightarrow F$ should now result in three different mappings. $N \longrightarrow F_r$ maps the place name to the *raw* footprint consisting of the corresponding point cloud. $N \longrightarrow F_p$ maps to the *polygon* which approximates the region occupied by the point cloud. $N \longrightarrow F_c$ finally maps a place name to the footprint's *centroid*, i.e., to a single coordinate pair as returned by conventional gazetteers. The centroid is the mean of *all* coordinate pairs in the point cloud and is thus specifically (and intentionally) biased towards areas that contain high numbers of geotags. F_c can thus be regarded as the point of interest best representing a place name, based on the number and location of corresponding geotags.

While the derivation of the gazetteer entries from geotags allows for enhanced functionality in the mapping from place name to footprint, the mapping to place type $N \longrightarrow T$ remains unchanged. The experimental setup presented in Section 4 leaves the place type unspecified. Potential combinations with linguistic approaches [21] as sketched in Figure 1, however, would allow for a semi-automatic classification of the gazetteer entries based on a predefined typing scheme. This scheme could be adopted from existing gazetteers. Due to the limited reliability of any data coming from such collaborative platforms, such an approach would at least require quality control mechanisms. A fully automatic *strong* typing of place names with such bottom-up approach is clearly not feasible here. While this is out of scope for this paper, the grouping of a resource's tags into place names, place types and other tags does appear feasible. Moreover, it stands to reason whether such a tag-based typing is a more practical approach for a community-driven gazetteer [32].

4 Workflow and Algorithm

This section describes the crawling approach implemented in our prototype. The different aspects of the resources that play a role in the filtering process are discussed.

4.1 Crawling Approach

A reliable extraction of geographic footprints requires a sufficiently large number of geotagged resources. We have limited ourselves to photos as resources for various reasons. People sharing their creations on the web want others' recognition.

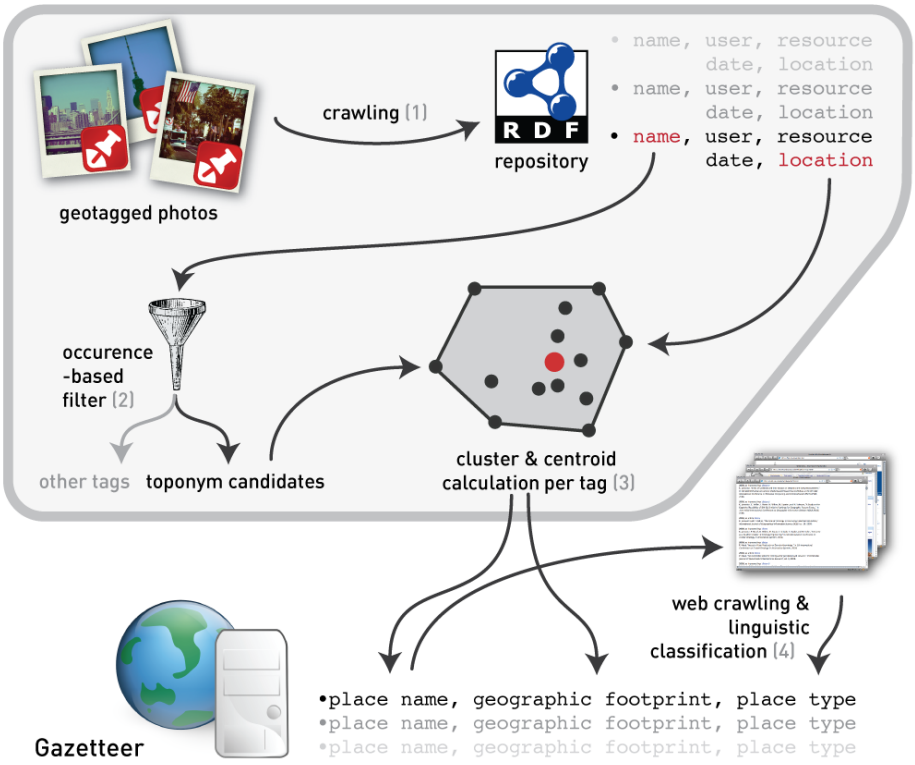

Fig. 1. Geotagged photos are crawled from the web (1) and fed into an RDF triple store. The tags are filtered based on occurrences to retrieve a subset of toponyms (2). For each place name, regions and centroids are calculated (3). Finally, every place name is categorized using linguistic classification (4). The part outlined in grey has been implemented for this paper (adapted from [7]).

Community-based web sites take this aspect into account by ranking the photos by popularity, which relies on the *findability* of the photos. Photo-sharing web sites all provide various means to find a photo: one can use a keyword-based search engine, browse a map with overlaid pictures, browse pictures by date, and so on. Users spent a considerable amount of time to annotate the pictures to cover all these aspects. Since every photo is implicitly located, assigning an explicit location by linking the photo to a point on a base map is a common annotation procedure. Accordingly, digital photos do not only carry detailed metadata in their Exif tags, they are also exceptionally well described by their creators. The last and most important reason to consider only photos as resource for extracting the spatial footprints of place names is the abundant availability. It is therefore reasonable to assume that the crawling yields a large enough sample of geotagged resources to achieve a significant result.

The crawling algorithm is conceptually straight-forward. Starting from a specific tag, the algorithm requests all geotagged resources which have been annotated with this tag. All three services used for our study provide this functionality through their APIs. For every tag attached to a retrieved photo, we store a separate complete geotag tuple (L, U, C, I, T) in our RDF triple store. In the next step, the conditions detailed in section 3.1 are applied to filter out tags which we have identified as not important. The resulting set of geotag tuples is taken as input for the clustering method described in the following.

4.2 Geotag Extraction Algorithm

A place name either refers to one unique place (e.g. *Kilimajaro*) or to multiple regions (e.g. the districts *Soho* in London and New York). The geotag tuples resulting from the crawling algorithm are used to identify clusters of high point-density. We consider the point cloud (explained in Section 3.3) as geographic footprint for the label L_i if many people used this keyword to annotate their photos taken nearby. Such clusters can have any shape, they are not necessarily concave and can contain holes. Point clouds derived from geotags are not equally distributed over space, but have some tendency to follow structures like trails or streets. In [10] the Delaunay triangulation has been identified as candidate algorithm to find clusters within point clouds. This method is not restricted to places with certain geometries. It computes the smallest possible triangle between three adjacent points; each point is connected to its nearest neighbors by an edge. A Delaunay triangulation for the tag *Soho* in New York is depicted in Figure 2. In order to split the graph of points and edges into clusters of high density (short edges), we remove all edges longer than a given threshold. If adjacent, remaining triangles are merged into one or more polygons. They represent F_p, the polygonal geographic footprint of the gazetteer's place name N.

A more advanced way to extract polygonal footprints from single locations is the Alpha Shape [33,34], which has also been used to generate the Flickr shape files[7]. For reasons of simplicity, we sticked to a Delaunay triangulation for this experiment. The next section shows that even with such a comparably simple clustering approach one can already obtain usable results.

5 Experimental Results and Evaluation

This section presents the results obtained by our prototype implementation. The results are discussed and compared to those obtained from conventional gazetteers.

5.1 Soho, Camino de Santiago and Kilimanjaro

We retrieved geotagged photos annotated with *Soho*, *Camino de Santiago* and *Kilimanjaro*. These three place names were chosen because they represent different geometries: Soho as a city district represents polygonal real-world features

[7] See http://code.flickr.com/blog/2008/10/30/.

Fig. 2. Cluster graph after the Delaunay triangulation for the place name *Soho*. The screen shot shows the clustering result depending on the edge length threshold: A small value results in several small clusters shown in blue. When the threshold increases, the fragments starts to join to the large black cluster.

up to a few kilometers in diameter. Moreover, we chose this example because there is not "the one" Soho, but both districts in London and New York can be regarded as equally well-known. Camino de Santiago refers to a number of pilgrimage routes leading to the Cathedral of Santiago de Compostela[8] in northwestern Spain. It usually refers to *Camino Francés*, the medieval route along Jaca, Pamplona, Estella, Burgos and León, but it is also used for a number of other ways to Santiago de Compostela across Europe and is thus a prime example of an ambiguous linear real-world feature. The third example, Kilimanjaro, is an example of a large-scale natural feature that can be seen (and hence shot) from far, but is hard to reach. Using this example, we want to investigate how well our approach is apt to derive useful results for such features.

Table 1. Figures on the RDF repository used for this study. The numbers include a negligible number of entries added during the testing phase.

Geotag Tupels	**Filtered Geotag Tupels**	**Unique Names**	**Filtered Unique Names**	**Resources**	**Users**
560,834	471,393	9,917	2,035	10,603	1,103

Table 1 gives an overview of the number of resources and tags obtained by crawling the three photo sharing websites for the three given examples. Only around 15 percent of tuples were removed during the filtering process, the ratio

[8] Tradition has it that the cathedral contains apostle Saint James the Great's gravesite.

of ∼0.84 is surprisingly high. The ratio from filtered to unfiltered unique names on the other hand is ∼0.21; this shows that our filtering approach identified almost 80% of the names as irrelevant since they were used by only one user. The difference between the two ratios means that the remaining 20% of filtered uniques names appear in 80% of all geotag tuples. Our rather simple approach of not further considering tags that only occur once thus proves very effective. Most tags are noise, but those which remain are used and accepted by many users. Table 2 contains the specific numbers per place name.

For Soho, the two biggest clusters emerge as expected in central London and in New York (see Figure 3). Apart from these two main clusters, a number of smaller clusters appear at different locations around the world. An analysis of the corresponding resources showed that most of them correspond to smaller places called Soho, thus representing valid gazetteer entries. The small outlying clusters south of the main cluster in Figure 3, however, are clearly no meaningful results. Such outliers occur frequently when users tag whole photo sets with the name of the place where *most* of them were taken. This inevitably tags some photos with the wrong place name and will require an improved filtering approach.

Table 2. Figures on the three case studies. The last column indicates the distance from the cluster's centroid to the corresponding footprint in GeoNames (a: London, b: New York).

Place name	**Geotag Tuples**	**Resources**	**Users**	**Dates**	**Distance**
Soho	11916	3124	446	3087	0.26^a/ 0.16^b km
Camino de Santiago	5132	1304	75	1255	285.3 km
Kilimanjaro	2536	825	72	808	3.7 km

For Camino de Santiago, the generated clusters give a good impression of the main trail to the Cathedral of Santiago de Compostela (see Figure 4). One apparent problem here is that the clustering algorithm splits up the route into distinct segments. Future research should focus on the development of "intelligent" clustering approaches that take the shape of the cluster into account, in order to enable a more reliable clustering.

For Kilimanjaro, the emerging clusters (see Figure 5) expose the main problem with an approach based on tagged and geolocated photos: users often do not tag the picture with the place name of the location where the picture was taken, but with the name of real-world feature *shown* in the picture. This becomes especially apparent for very large real-world features, as in this example. Several smaller clusters expose the high number of pictures taken at these locations, which apparently offer a good view on Mount Kibo, the highest peak of the Kilimanjaro massif. Future work needs to investigate how clusters referring to such real-world features can be detected, for example, by identifying ring-shaped clusters such as the one in Figure 5.

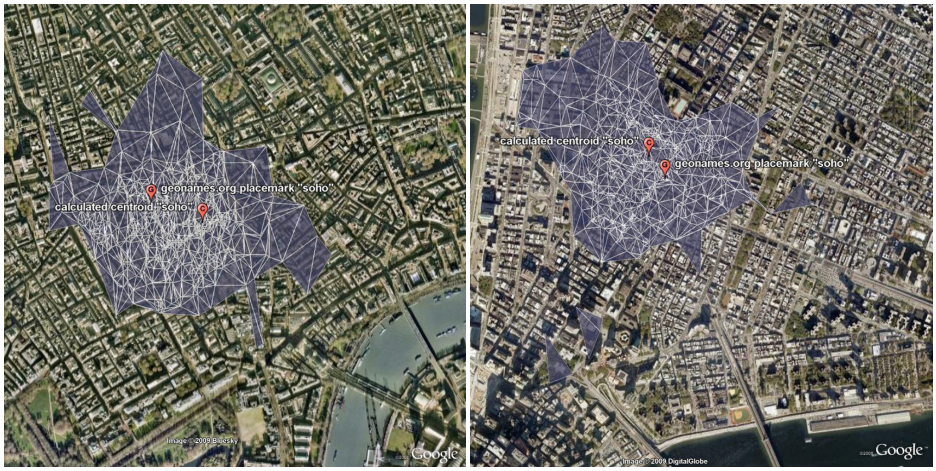

Fig. 3. The clusters generated for *Soho*. The left screen shot shows the cluster in London, the right one shows the cluster in Manhattan, New York.

5.2 Geographic Footprints

The footprints extracted by our approach provide additional useful information to the point-based footprints provided by conventional gazetteers. For comparison with GeoNames, we also computed the corresponding centroid as the mean of all coordinates in every cluster (or cluster group, as for Kilimanjaro). This centroid points to what can be described as a named cluster's *group-cognitive centre*. In contrast to the geometric centre point, it gives an estimate of the common point of interest of users providing the photos retrieved in the crawling step. In the following, we discuss the extracted footprints and how the group-cognitive centre and the geometric centre point differ for our three examples.

For Soho and Kilimanjaro, the distance between the GeoNames footprint and the centroid of our cluster is comparably small, given the respective scale of the cluster (and the size of the corresponding real-world feature). The footprint for Soho, London, in GeoNames is about 260m away from the centroid of our cluster. The cluster itself represents the common notion of Soho very well[9], although it extends across Oxford Street in the north, which is usually taken as Soho's northern border. The same applies to the eastern extension of the cluster; the southern and western extension match the common notion of Soho very well. Similar observations can be made for Soho, New York: The area that is commonly referred to as Soho[10] is completely covered, but the cluster exceeds the actual area in all four directions. This exceeding problem can probably be addressed by adjusting the cutoff length during triangulation and fetching more input data. The centroid of the cluster is only 160m away from the footprint of the

[9] See http://en.wikipedia.org/wiki/Soho#Streets for comparison.
[10] See http://en.wikipedia.org/wiki/SoHo#Geography.

Fig. 4. The clusters generated for *Camino de Santiago* give a good impression of the trail of the route.

corresponding GeoNames entry. The clusters generated for Camino de Santiago stretch very well along the actual trail of the route, despite the gaps discussed above. The calculation of the centroid shows that it is in most cases meaningless to represent linear real-world features by points. While the centroid represents a mean value for all coordinates in the clusters, the footprint from GeoNames is located at one end of the route. Selecting the destination of the pilgrimage trail as footprint certainly makes sense in this case (the coordinate refers to Santiago de Compostela), however, this selection will be completely arbitrary for linear features that lack such a clear destination (such as most roads). For Kilimanjaro, the clusters represent the areas with a *view* on the Kilimanjaro's highest peak, rather than the mountain itself (due to the problems discussed above). This also causes a distance of almost 4 km of the clusters' centroid to the GeoNames footprint, which is nevertheless still within an acceptable range given the size of the real-world feature.

6 Conclusions

This section summarizes the paper and points to different applications of the approach presented in this paper, as well as directions for future work.

6.1 Discussion

In this paper, we have presented an experiment to test the feasibility of the idea to build a gazetteer completely from geotagged photos crawled from the web. We have introduced the theoretical foundations to capture the emergent semantics of geographic information extracted from geotagged resources on the web. A theoretically sound definition of a *geotag* has been introduced and related to the classical definition of a gazetteer. Using the implementation which clustered and

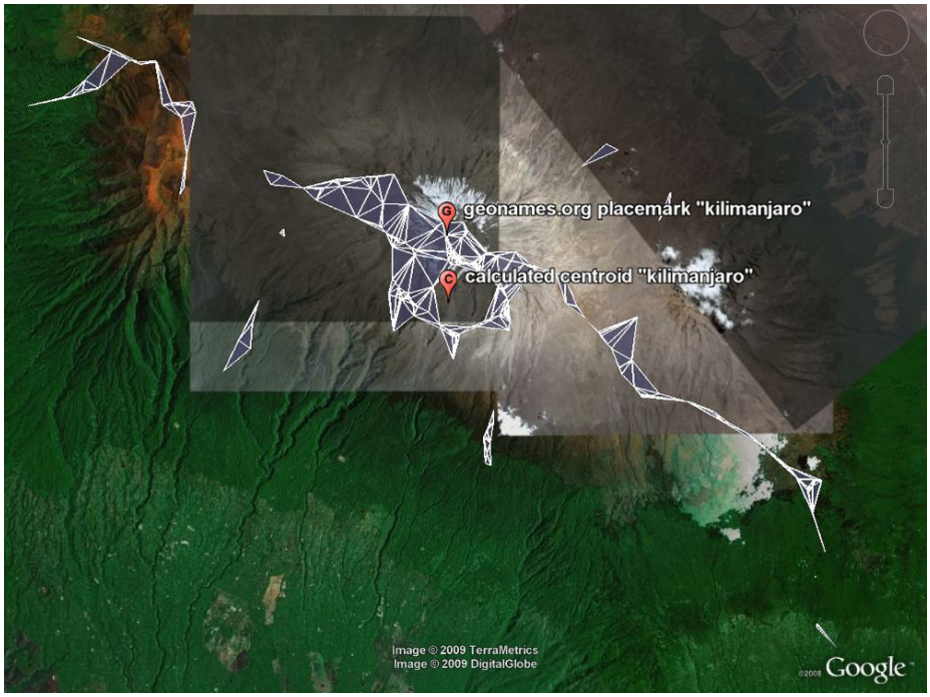

Fig. 5. The clusters generated for *Kilimanjaro* are distributed over a large area and show the problem of photos tagged with with the names of features shown in the pictures, although they were taken from far away.

filtered geotags of photos, we have demonstrated how the geographic footprint for a given place name can be derived.

The results of our queries for *Soho*, *Camino de Santiago* and *Kilimanjaro* showed that it is possible to derive meaningful geographic footprints from geotagged content, even with comparably simple clustering approaches. Both the footprints as well as their centroids shed a different light on named places than conventional gazetteers. As pointed out in [22], every gazetteer extracted from online information can only be as good as the information it builds on. However, our experiment has demonstrated that useful results can already be obtained with very straight-forward means to extract a group-cognitive perspective [12] on place names. Hence, we do not propose to replace existing gazetteers by our approach, but to complement them within a gazetteer infrastructure [7,16]. Further improvements can be expected from implementing models of trust in the harvesting process, which would allow for an estimation of the quality of the geotags used for clustering [7,23].

From a visual inspection, the generated regions were judged to be plausible representations of the place names' geographic footprints. Particularly, the algorithm showed the capability to recognize different places carrying the same name,

as shown in the *Soho* example. Moreover, the filtering algorithm has successfully sorted the crawled tags into toponyms and other tags based on the notion of significant occurrences. The example of *Kilimanjaro* has shown that very large real-world features are problematic for our approach, since they often appear in the context of photos that show them, but that were taken far away from the actual feature. Evidently, the results could be improved by more sophisticated crawling, filtering and clustering approaches.

6.2 Applications

While the crawling approach presented in this paper has been developed with the recursive generation of a bottom-up gazetteer in mind, the underlying algorithms are also potentially useful in a number of other applications. The user component, for example, could be used to derive communities and their vocabulary by analyzing how groups of users tag certain real-world features. The temporal component has only been used to identify occurrences and to filter events that might corrupt the place name recognition. Instead of treating these filtered events as noise, however, one could also imagine an application that specifically looks for such events based on temporal clusters. This would enable an automatic calculation of geographic footprints for such events, which could eventually be merged into event gazetteers [35,36].

The fact that every resource carries a time stamp and a user's name can also be used to extract individual space-time prisms [37,38]. This may provide insight into real-world social interactions between the users of photo sharing platforms, such as "who travelled together" or "who went to this party". The implications for privacy, however, are obvious and would require a careful consideration of ethical issues. From this perspective, the photo sharing platforms used in this paper might require more fine-grained mechanisms to give their users control over what information they want to reveal to whom. One method to prevent automatic generation of such profiles would be to allow users to exclude specific metadata (or combinations of them) from access through the respective APIs.

6.3 Future Work

The next step in this research will be the combination of the filtering and clustering algorithm presented in this paper with linguistic web crawling approaches. This would facilitate to go beyond place names and their geographic footprints and also extract the corresponding place type, as demonstrated by Uryupina [21]. It is, however, unlikely that it will also be possible to extract a *strong* place typing from user tags. While straightforward types such as *city*, *street* or *river* may still be found frequently enough in the tags for a reliable extraction, it is unlikely that a user tags a picture taken in Soho with *section of populated place* – the associated feature class (i.e., place type) in GeoNames. However, same as for footprints and centroids, such a bottom-up typing scheme would reflect place types used in common language, as opposed to the often somewhat artificial administrative place types used in current gazetteers. This bottom-up

approach should also allow for a more flexible categorization that does not force every named place into exactly one category [32] in order to fully capture the emergent semantics of collections of geotagged content. We also plan to extend the existing implementation to take the temporal nature of geotags into account. This eventually results in the identification not only of place names, but also of names of events and processes with a spatial character.

Acknowledgments

This research has been partly funded by the SimCat project (DFG Ra1062/2-1 and DFG Ja1709/2-2, see http://sim-dl.sourceforge.net) and the GDI-Grid project (BMBF 01IG07012, see http://www.gdi-grid.de). Figure 1 contains geotag icons under a Creative Commons license from http://geotagicons.com.

References

1. Jones, C.B., Purves, R.S., Clough, P.D., Joho, H.: Modelling vague places with knowledge from the web. International Journal of Geographical Information Science **22**(10) (2008) 1045–1065
2. Larson, R.R.: Geographic information retrieval and spatial browsing. GIS and Libraries: Patrons, Maps and Spatial Information (April 1996) 81–124
3. Goodchild, M.F.: Citizens as voluntary sensors: Spatial data infrastructure in the world of web 2.0. International Journal of Spatial Data Infrastructures Research **2** (2007) 24–32
4. Bennett, B., Mallenby, D., Third, A.: An ontology for grounding vague geographic terms. In Eschenbach, C., Gruninger, M., eds.: Proceedings of the 5th International Conference on Formal Ontology in Information Systems (FOIS-08), IOS Press (2008)
5. Henrich, A., Lüdecke, V.: Determining geographic representations for arbitrary concepts at query time. In: LOCWEB '08: Proceedings of the first international workshop on Location and the web, New York, NY, USA, ACM (2008) 17–24
6. McConchie, A.: The great pop vs. soda controversy, available from http://popvssoda.com (last visited agust 1^{st}, 2009) (2002)
7. Keßler, C., Janowicz, K., Bishr, M.: An agenda for the next generation gazetteer: Geographic information contribution and retrieval. In: ACM GIS '09, November 4–6, 2009. Seattle, WA, USA, ACM (2009)
8. Wilske, F.: Approximation of neighborhood boundaries using collaborative tagging systems. In Pebesma, E., Bishr, M., Bartoschek, T., eds.: GI-Days 2008. ifgiPrints 32 (2008) 179–187
9. Guo, Q., Liu, Y., Wieczorek, J.: Georeferencing locality descriptions and computing associated uncertainty using a probabilistic approach. International Journal of Geographical Information Science **22**(10) (2008) 1067–1090
10. Heuer, J.T., Dupke, S.: Towards a spatial search engine using geotags. In Probst, F., Keßler, C., eds.: GI-Days 2007 – Young Researchers Conference. ifgiPrints 30 (2007) 199–204
11. Aberer, K., Mauroux, P.C., Ouksel, A.M., Catarci, T., Hacid, M.S., Illarramendi, A., Kashyap, V., Mecella, M., Mena, E., Neuhold, E.J., Et: Emergent semantics principles and issues. In: Database Systems for Advances Applications (DASFAA 2004), Proceedings, Springer (March 2004) 25–38

12. Stahl, G.: Group Cognition: Computer Support for Building Collaborative Knowledge (Acting with Technology). MIT Press (2006)
13. Raubal, M.: Cognitive engineering for geographic information science. Geography Compass **3**(3) (2009) 1087–1104
14. Surowiecki, J.: The Wisdom of Crowds. Anchor (2005)
15. Schlieder, C.: Modeling collaborative semantics with a geographic recommender. In: Advances in Conceptual Modeling – Foundations and Applications. Volume 4802 of Lecture Notes in Computer Science. Springer (2007) 338–347
16. Janowicz, K., Keßler, C.: The role of ontology in improving gazetteer interaction. International Journal of Geographical Information Science **22**(10) (2008) 1129–1157
17. Hill, L.L.: Georeferencing: The Geographic Associations of Information (Digital Libraries and Electronic Publishing). The MIT Press (2006)
18. Casati, R., Varzi, A.C.: Parts and Places. The Structures of Spatial Representation. MIT Press, Cambridge and London (1999)
19. Goodchild, M.F., Hill, L.L.: Introduction to digital gazetteer research. International Journal of Geographical Information Science **22**(10) (2008) 1039–1044
20. Hastings, J.T.: Automated conflation of digital gazetteer data. International Journal of Geographical Information Science **22** (October 2008) 1109–1127
21. Uryupina, O.: Semi-supervised learning of geographical gazetteers from the internet. In: Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references, Morristown, NJ, USA, Association for Computational Linguistics (2003) 18–25
22. Goldberg, D.W., Wilson, J.P., Knoblock, C.A.: Extracting geographic features from the internet to automatically build detailed regional gazetteers. International Journal of Geographical Information Science **23**(1) (2009) 93–128
23. Bishr, M., Kuhn, W.: Geospatial information bottom-up: A matter of trust and semantics. In Fabrikant, S., Wachowicz, M., eds.: The European Information Society – Leading the Way with Geo-information (Proceedings of AGILE 2007, Aalborg, DK). Springer Lecture Notes in Geoinformation and Cartography (2007) 365–387
24. Guszlev, A., Lukács, L.: Folksonomy & landscape regions. In Probst, F., Keßler, C., eds.: GI-Days 2007 – Young Researchers Conference. ifgiPrints 30 (2007) 193–197
25. Gruber, T.: Ontology of folksonomy: A mash-up of apples and oranges. International Journal on Semantic Web & Information Systems **3** (2007 – originally published at http://tomgruber.org/writing/ontology-of-folksonomy.htm in Nov. 2005)
26. Frank, A.: Ontology for spatio-temporal databases. In Koubarakis, M., Sellis, T., Frank, A.U., Grumbach, S., Güting, R.H., Jensen, C.S., Lorentzos, N., Manolopoulos, Y., Nardelli, E., Pernici, B., Schek, H.J., Scholl, M., Theodoulidis, B., Tryfona, N., eds.: Spatio-Temporal Databases. Lecture Notes in Computer Science. Springer (2003) 9–77
27. Goodchild, M.F.: Geographical data modeling. Computational Geosciences **18**(4) (1992) 401–408
28. Saeed, J.I.: Semantics (Introducing Linguistics). Wiley-Blackwell (2003)
29. Searle, J.R.: Proper names. Mind **67**(266) (1958) 166–173
30. Codd, E.F.: A relational model of data for large shared data banks. Communications of the ACM **13**(6) (1970) 377–387
31. O'connor, M., Tu, S., Nyulas, C., Das, A., Musen, M.: Querying the semantic web with swrl. (2007) 155–159
32. Shirky, C.: Ontology is overrated – categories, links, and tags. Essay available from http://shirky.com/writings/ontology_overrated.html (2005)

33. Edelsbrunner, H., Kirkpatrick, D., Seidel, R.: On the shape of a set of points in the plane. Information Theory, IEEE Transactions on **29**(4) (1983) 551–559
34. Edelsbrunner, H., Mücke, E.: Three-dimensional alpha shapes. ACM Transactions on Graphics **13**(1) (1994) 43–72
35. Allen, R.: A query interface for an event gazetteer. In: Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries. (2004) 72–73
36. Mostern, R., Johnson, I.: From named place to naming event: creating gazetteers for history. International Journal of Geographical Information Science **22**(10) (2008) 1091–1108
37. Hägerstrand, T.: What about people in regional science? Papers in Regional Science **24**(1) (1970) 6–21
38. Miller, H.J.: A measurement theory for time geography. Geographical Analysis **37** (2005) 17–45