Towards a crowdsourced supervision of the analysis of user-generated geographic content: Engaging citizens in discovering urban places

Frank O. Ostermann*, Gustavo A. García-Chapeton, Menno-Jan Kraak, Raul Zurita-Milla

Department of Geo-Information Processing, Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands

* Corresponding Author e-mail: f.o.ostermann@utwente.nl

This paper presents a conceptual model and a proof-of-concept implementation of a novel approach to engage citizens in supervising the analysis of user-generated geographic content (UGGC). For example, UGGC allows insights into the (re-) production of urban spaces, promising new options for pro-active urban planning, citizen participation, and local empowerment. However, the unknown quality and high volume of UGGC require advanced filtering and classification procedures to be able to extract new knowledge and actionable intelligence. The complexity of some tasks and the increasing volume of UGGC often restrict the role of citizens to data providers. We argue that citizen should and can play a decisive role in the parameterization and training of deeper computational analysis of UGGC streams. The challenge is how to present geographical analysis problems to a crowd of human supervisors, and how to elicit responses and feed those back into the workflow. We propose a hybrid processing approach, which maps geographical problems into data mining and machine-learning tasks, presents analysis results to human supervisors, and uses the responses to improve the machine-learning and data mining. For the pilot study, we adapt an approach to find semantically distinct places in UGGC. The human supervisors rate the clustering of potentially similar geo-located photographs from the platform Flickr, and thereby help parametrize both the data mining of geospatial clusters, as well as the classification of similar images based on several ancillary attributes.

Keywords: crowdsourcing, citizen science, machine learning, data mining, user-generated geographic content, Flickr, urban places, hybrid processing

1 Introduction

1.1 Problem Statement and Motivation

This paper explores a novel approach for crowdsourcing the supervision of computational analysis of user-generated geographic content to citizens and potential stakeholders in the study objectives and outcomes. This approach takes advantage of the rapid growth of available sensor data and user content. With many human activities relating to geographic space, this user-generated content (UGC) also contains rich information about geographic features or about events occurring in a specific time and place (Caquard, 2014; Graham and Shelton, 2013), allowing us to observe the development of urban imaginaries (Kelley, 2013), and influencing the (re-)production of space (Thrift and French, 2002). The low technological entrance barrier, wide reach and bi-directional communication channels open up new opportunities for citizens to engage in participative planning processes (D'Hondt et al., 2013) or to challenge authoritative knowledge production (Dodge and Kitchin, 2013). Citizens can participate in scientific projects more easily, leading to a surge in citizen science projects (Kullenberg and Kasperowski, 2016), and through their valuable traditional and local knowledge improve our understanding of processes that drive human-environment interaction (Bonney et al., 2009). There is a variety of terms used to describe the facets of geographic UGC, including *volunteered geographic information* (Goodchild, 2007), *ambient geographic information* (Stefanidis et al., 2013), and *contributed geographic information* (Harvey, 2013). In this paper, we focus on content from geosocial media and refer to it as user-generated geographic content, or UGGC (Craglia et al., 2012).

Although UGGC has proven its utility for a variety of tasks and purposes (Fast and Rinner, 2014; Garcia-Martí et al., 2016; Granell and Ostermann, 2016; Haworth, 2016), several characteristics of UGGC have a negative impact on its fitness-for-use. These include high semantic and syntactic heterogeneity, and unknown provenance and production parameters (Ostermann et al., 2015).

A common response to these challenges has been the crowdsourcing of curation tasks (Sui et al., 2012): Human volunteers curate incoming information by checking its accuracy, assigning labels, and prioritizing it for further processing. Despite encouraging results, this approach lacks quality (control) and reproducibility (Camponovo and

Freundschuh, 2014), guaranteed sustainability, and efficient scaling up (Morrow et al., 2011).

Another approach is to employ data mining (DM) and machine learning (ML) techniques to select, filter, classify, and enrich UGGC. Already more than a decade ago, Gahegan (2003) identified ML as a promising approach to many (geographic) analysis tasks, emphasizing its potential for inductively generating new knowledge. Since then, research and practice revealed at least three main challenges: First, dependency on data quality for unsupervised DM and ML (Kanevski et al., 2008); second, overfitting of the learning model (Butler, 2013); and third, training costs for a diversity of contexts and tasks (Ostermann and Spinsanti, 2012).

1.2 Research Objectives and Paper Structure

The overarching objective of this paper is to develop and test mechanisms that address the various challenges discussed in the previous section, by combining human and computational analysis of UGGC. We envision that such a hybrid processing approach can improve the quality assessment and enrichment of near real-time UGGC streams, exploiting contextual, local, or traditional knowledge from the human supervisors to increase UGGC's fitness-for-use. For any scientific analysis, such an approach can contribute to ensure that any results are not only statistically significant, but also meaningful. This paper investigates three main research objectives, all with a focus on UGGC streams:

1. What are the characteristics of a hybrid processing workflow, and which of the identified challenges can it address?

2. How can we identify which geospatial analysis tasks and methods are suitable for a hybrid processing workflow, and implement them in a reproducible manner?

3. What are characteristics of a feasible system architecture for implementing such a hybrid processing workflow?

We approached these questions in two distinct steps: First, based on a review of related work (section 2), we address research questions 1 and 2 by developing a conceptual model and formulating general principles that facilitate the transfer to other system set ups and case studies (section 3). Second, to address research question 3, we test a

proof-of-concept prototype to demonstrate the approach (section 4) by attempting to discover distinct urban places from meta-data of Flickr images. The paper continues with a discussion (section 5) of the approach and its results, before drawing conclusions and presenting an outlook (section 6).

2 Related Work on Processing UGGC

Many projects especially from the crisis management domain have tried to process UGGC using a massively collaborative approach, relying on asynchronous distributed interaction to achieve greater reach and flexibility. Volunteers would, for example, translate and geolocate text messages such as SMS or Tweets using web interfaces (Liu, 2014). Another example is a map-based approach of the Humanitarian OSM Team, where volunteers create new geographic information from satellite or aerial imagery (Kelso, 2010), often gathering in mapping parties for face-to-face collaboration. An often cited example for this approach is the Haiti 2010 earthquake (Meier, 2013), after which the crisis mapper community experienced a boost. The crisis mappers' Stand-by-task-force[1] developed standardized operating procedures and refined them during practices. Important governmental and on-governmental organizations embraced this new digital humanitarianism, resulting in or contributing to valuable tools such as the Humanitarian Exchange Language (Kessler and Hendrix, 2015).

Computational methods offer an alternative approach to data curation. Briefly, DM is about discovering new patterns in data. It is an inductive approach, which aims to describe new laws and generate knowledge by looking at the data, and observing correlations and patterns. DM is mostly unsupervised, meaning that the algorithms extract patterns and parameters without significant human input (although the choice of parameters is crucial). ML on the other hand is a more deductive approach, which relies on human supervision to classify information, and uses this annotated information as training set to train a classifier for processing new, unlabelled information. It tries to apply existing knowledge reliably to new information. Most of the geospatial ML has been on physical geographic phenomena (Kanevski et al., 2009). Recent works on anthropogenic phenom-

[1] http://blog.standbytaskforce.com/

ena include disambiguation of place references (Santos et al., 2015), classification of micro-blogging messages (Imran et al., 2013), urban sensing (Kaiser and Pozdnoukhov, 2013), gamification (Barrington et al., 2012), and detection of information about forest fires (Spinsanti and Ostermann, 2013).

There have been few attempts to combine human and machine computation for processing Big Data in general and UGGC in particular. For non-geographic data, the focus has been on large-scale classification efforts involving many classes and many human annotators. The lessons learnt include strategies to address the complexity of the data (Sun et al., 2014), methods to organize and motivate participants (Burger et al., 2011), and the decision on when to stop labelling (Kamar et al., 2012).

The AIDR system (Imran et al., 2014) focuses on adaptive aggregation and filtering of Twitter, integrating crowd-sourced labelling to learn rules to filter and classify social media information. It is open source and allows near real-time processing. However, currently AIDR relies on a single source (Twitter) and focuses entirely on the content, ignoring geographic semantics.

The Twitcident (Abel et al., 2012) aggregates and filters social media around events extracted from emergency broadcasting services. It semantically enriches the incoming information and links it with other external information. However, location only seems to influence the filtering of the information and not the assessment. It seems that a recent extension (Crowdsense) enables it to use several social media sources. No source could be found.

The GeoCONAVI system (Spinsanti and Ostermann, 2013) is capable of detecting past forest fire events. In contrast to the previous systems, it uses multiple sources (Tweets, Flickr image meta-data), and exploits geographic semantics by enriching the UGGC with geographic context and spatio-temporally clustering it. The processing is done in high-frequency batch processing. The content classification employs decision trees trained on an event-specific annotated data set. Case study results (Ostermann and Spinsanti, 2012) show a low false positive rate (high specificity), and a low false negative rate (high sensitivity). It does not examine the source, nor does has it been adapted to other event types.

3 Developing a Framework for Hybrid Processing

3.1 Introducing Hybrid Elements to a UGGC Processing Workflow

To recall, the main challenges of UGGC are its high semantic and syntactic heterogeneity, and unknown provenance and production parameters. In principle, these can be solved with human curation or machine computation, given enough resources. However, in practice either of these approaches faces serious challenges of its own. Human curation is subject to unknown inaccuracy and a lack of reproducibility, scalability, and sustainability. We argue that these can be addressed by using complementary computational and crowdsourcing approaches, which in turn face the complementary challenges of dependency on data quality for unsupervised ML, overfitting of the learning model, and diversity of contexts and tasks. We further argue that these can be addressed in turn with human curation (compare Figure 1).

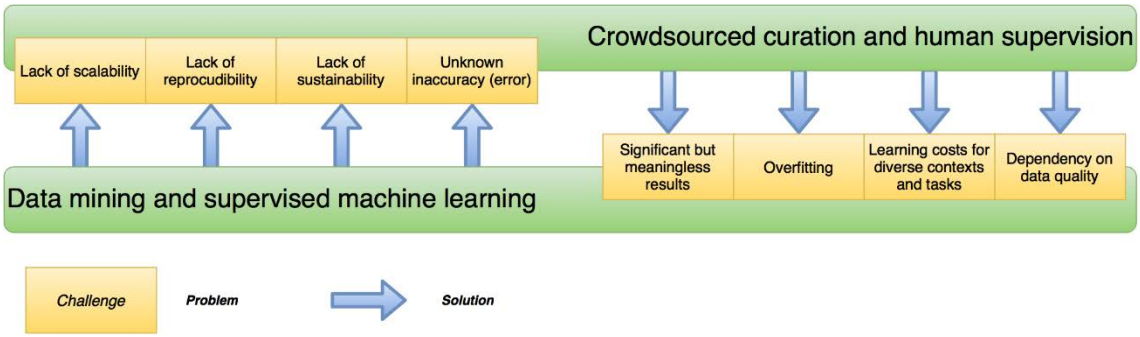

Figure 1 Hybrid approach addressing challenges of UGGC processing

A typical UGGC processing workflow, of which Figure 2 shows an abstracted simplified representation, has several phases in which DM and ML techniques can play an important role. After the collection of potentially relevant UGGC from the public APIs, supervised learning algorithms classify the content, in order to reduce the noise in subsequent analysis steps. If the UGGC content or meta-data contains toponyms (place names), these require geo-coding using natural language processing procedures (named entity recognition, disambiguation). Multiple found toponyms might be synthesized into a single geospatial footprint to represent the UGGC's geographic coverage. A second classifier checks the validity of such derived results. After georeferencing, it is possible to enrich and contextualize the UGGC with additional relevant information from authoritative or non-authoritative sources. Next, the individual items are clustered to detect events or

patterns in space and time. Secondary aims are to further reduce redundant content, and reinforce the credibility of confirmed content. Again, this clustering can profit greatly from human supervision to parametrize the clustering algorithm, and filter out any irrelevant or meaningless clusters. The last step aims to provide an integrative assessment of credibility and relevance, and to present actionable intelligence that for example contributes to increasing situational awareness in crisis situations. Such a combined approach pays special attention to geographic semantics, first by examining what else is there, and second by asking volunteers whether the results make sense. In the case of time-critical applications such as disaster response, this processing has to happen in near real-time or at least in high-frequency micro-batches.

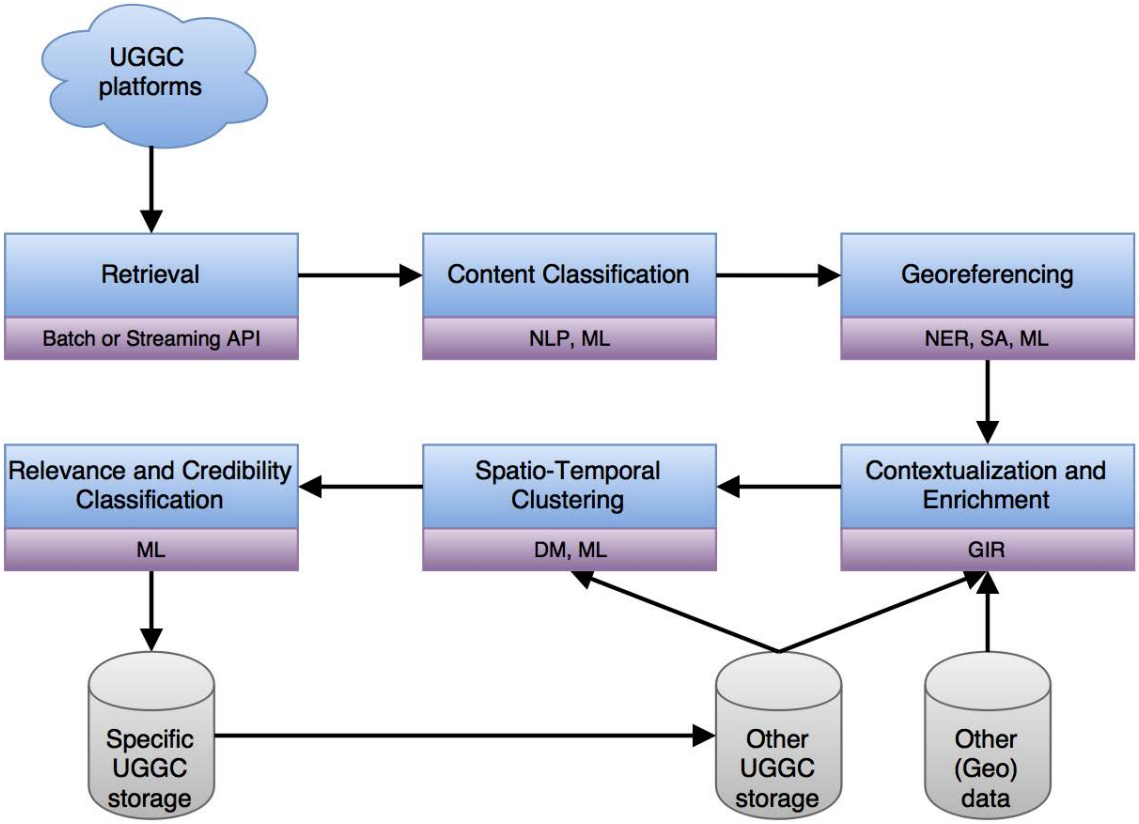

Figure 2 Typical UGGC processing workflow (API = Advanced Programming Interface, NLP = Natural Language Processing, NER = Named Entity Recognition, SA = Spatial Analysis, GIR = Geographic Information Retrieval, DM = Data Mining, ML = Machine Learning)

We can identify three main challenges for developing such a workflow of hybrid geoinformation processing: First, to link the specifics of geographic information with DM

and ML algorithms, and translate the annotation task to suitable questions for the human supervisors. Second, to develop an interface for human supervisors to annotate and label results and account for multiple annotations that contain noise. Third, to translate the learned models into nomothetic principles that can be applied in different contexts. Presenting an initial investigation, this paper focuses on the first and second challenges.

Following from the outlined UGGC workflow and the challenges it faces, we can identify four characteristics of an effective and efficient hybrid processing system:

1. Allowing (near) real-time processing to take advantage of UGGC streams
2. Multiple input sources to use geographic contextualization and geospatial semantics for assessing relevance and credibility of both data and human supervisor
3. Open source license to allow customization by any interested group of citizen developers
4. Modularity of components and formalization of supervision tasks to enable extensions and adaptations to new tasks and events

The following sub-section focus on the last criterion and addresses the second research question.

3.2 From Geospatial Analysis Tasks to Supervision and Validation Queries

A first step is to determine candidate tasks, for which crowdsourced supervision and validation could improve the performance. Since there are no comprehensive geospatial analysis task ontologies, this leaves us to develop our own representation. Geospatial analysis tasks exist at several scales, from low-level, simple tasks such as “create a buffer around all features in the data set,” to more complex, ill-defined tasks, e.g. “find all relevant data sets for assessing forest fire risk.” Given the scope of a comprehensive geospatial analysis task ontology, and the low adoption rate of many ontologies, we adopt a strategy of re-use (link to existing partial task ontologies whenever possible and feasible), bottom-up design (create what is needed), open world (allow for extensions), and standardization (rely on standards for implementing it).

Drawing on the GIS&T BoK (*Geographic Information Science and Technology Body of Knowledge*, 2006), candidates for suitable geospatial analysis tasks are: (point) patter analysis and clustering (AM5-1 to AM5-8), measures of spatial association (AM7-4 to AM7-7), autoregressive and geographically weighted regression models (AM9-2 to AM9-4), data mining and pattern recognition (AM10-2 to AM10-4), and geocomputation and classification (GC2-2 to GC2-6).

Grouping these, we make a fundamental distinction between hyper-parametrization of DM tasks and supervision of ML tasks. DM tasks in a UGGC processing context are mainly spatio-temporal clustering in order to discover unknown patterns in the data, but can also support the generation of hypotheses through abductive reasoning supported by geovisual exploration. ML tasks are often classifications, such as content classification (topicality), geo-coding (using the correct gazetteers), weighing the evidence, scoring the relevance and credibility of geo-information, or deciding on auxiliary (geo-) datasets. ML tasks could support (hyper-) parametrization and regression tasks

In order to crowdsource the supervision and validation in a hybrid processing work-flow, the tasks need to fulfill two criteria: First, it needs to deal with a human observable or recognizable geographic phenomenon; and second, it needs to provide an output that can be evaluated by lay persons and potentially profit from local knowledge that the an-notators possess.

Example supervision and validation tasks transformed into a query that can be asked through a (geo-) visual interface are:

- Content classification: "Is this [item] directly, indirectly, or not related to [topic]?"
- Geo-coding and toponym disambiguation: “Does this [item] talk about [location A] or [location B], or none, or both?”
- Deciding on auxiliary data sets: "Is [data set] related to [topic] and can help assess credibility of [item]?"
- Spatial footprint calculation for vague geographies: “Is this spatial footprint for [item] correct? If not, is it too large, too small, or wrong shaped, or wrong placed?”

- Spatio-temporal clustering: “Does this [item] belong to a cluster named [event] in [location]? If not, what’s wrong: Event, Location, or both?”

Some necessary core components of the model are thus [item], [topic], [location], [data set], where [data set] is a collection of [item]. In terms of Kuhn’s (2012) core geographic concepts, [item] corresponds to object, [location] obviously to location, and both neighborhood and event relate to what we want to discover in the collection of objects.

Figure 3 shows the most important elements for our bottom-up UGGC collaborative task ontology. It captures all necessary components as a starting point. A concrete example of the task ontology is provided in section 4.2.

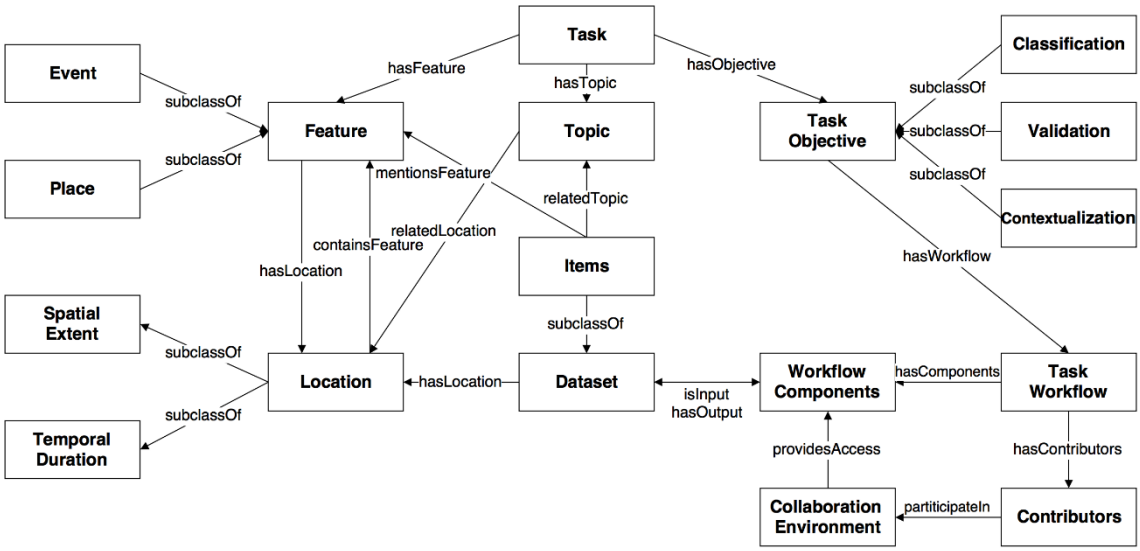

Figure 3 UML diagram showing the components of a geospatial task ontology

4 Pilot Study on Geotagged Photos

4.1 Study Area and Subject

We address the third research question by developing a pilot study. The primary aim of this pilot study is to demonstrate a feasible approach to hybrid geoinformation processing by applying the principles developed in the previous section. It prioritizes the use of well-tested and well-understood analysis methods over the optimization of results. The secondary aim is to investigate the relationship of UGGC and urban space and place. The motivation here is the need to develop representations of our geographic environment that take the multiplicity of perspectives into account, i.e. even for widely agreed-up upon

distinct urban places, we still expect considerable dissent on shape, size and theme of the places. From the various types of UGGC, shared image content on platforms such as Flickr, Panoramio, and Instagram has received a substantial amount of research interest, because photographs often possess a strong semantic link between content (image and descriptive text) and geographic location (Sigurbjörnsson and Van Zwol, 2008), and the annotations are potentially rich and personal. Flickr provides the most mature and accessible API. Flickr allows users to enter long titles and extensive descriptions for every image, and offers rich EXIF[2] metadata as well, including the orientation of the camera (if provided). Some examples of relevant studies include the delineation of vernacular place names (Hollenstein and Purves, 2010), the extraction of place tags from images based on location (Rattenbury and Naaman, 2009), and the detection of places (Van Canneyt et al., 2012), or the extraction of place semantics (Ostermann et al., 2015).

However, none of these have relied on a hybrid approach, and focused on a coarser granularity or did not validate all of the results. We adopt a bottom-up approach to mine the data set through geospatial clustering, instead of imposing a fixed grid over the study area similar to Feick and Robertson (2014). Further, we aim to validate the results to ensure that they represent meaningful places. To search for places, we look into the tags and descriptions of georeferenced photographs, using a controlled vocabulary of terms that describe activities, qualities, and elements of places (Purves et al., 2011; Tversky and Hemenway, 1983) to measure thematic similarity, and combine it with a spatial clustering looking for spatial proximity, and a classification to remove noise from the clustering. In terms of tasks identified in section 3.2, we focus on clustering (AM5), data mining (AM10), and classification (GC2). The prototype then presents the results of geographic analysis tasks to a small group of human study participants for a map-based supervision and validation.

Figure 4 (below) shows the pilot study adaptation using the previously developed task ontology (Figure 3) with unused components not shown, and instantiations of a class added to the class description box.

[2] EXIF (Exchangeable Image File Format) is a standard for metadata, used by digital cameras to record technical information of the camera's status when shooting a photograph.

As a study area, we chose the Greater London Area (GLA) because of its rich and diverse urban fabric, and abundance of large UGGC and open authoritative data sets. All these factors ensure the potential for deeper analysis.

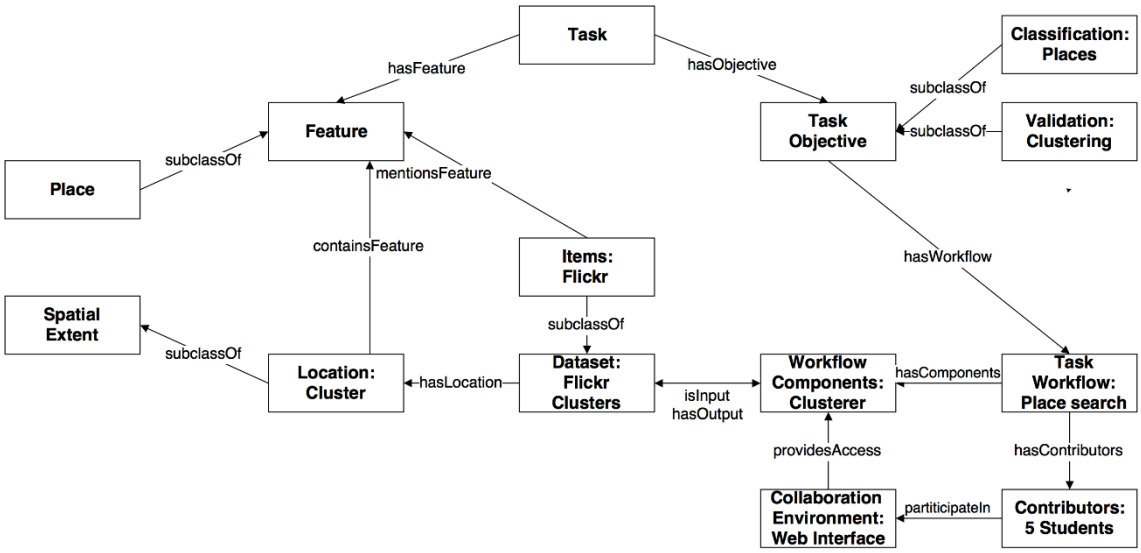

Figure 4 Task ontology adapted to pilot study

4.2 System Architecture and Set-up

The pilot study consists of eight main phases, which Figure 5 shows in an overview and which are explained in more detail below.

Phase 1. Collect Flickr photo meta-data

The initial step is to collect georeference Flickr image meta-data about the study area using the public API, and store the retrieved information in a PostGIS database.

Phase 2. Find place-related terms and build term vectors

Previous research (Purves et al., 2011) has established a list of English terms that are commonly used to describe Flickr images. They are grouped into the three sets of activities (107 terms, e.g. “party”, “football”, “exhibition”), elements (313 terms, e.g. “church”, “station”, “graffiti”), and qualities (161 terms, e.g. “dark”, “royal”, “woods”). For every image, the tags, title and descriptions were parsed to find any of these terms through lexical matching. We are aware that our approach might not find all tags if they

are misspelled or appear in composite words or expressions. More advanced NLP techniques can result in better recall, but the large number of images in our dataset compensates for this.

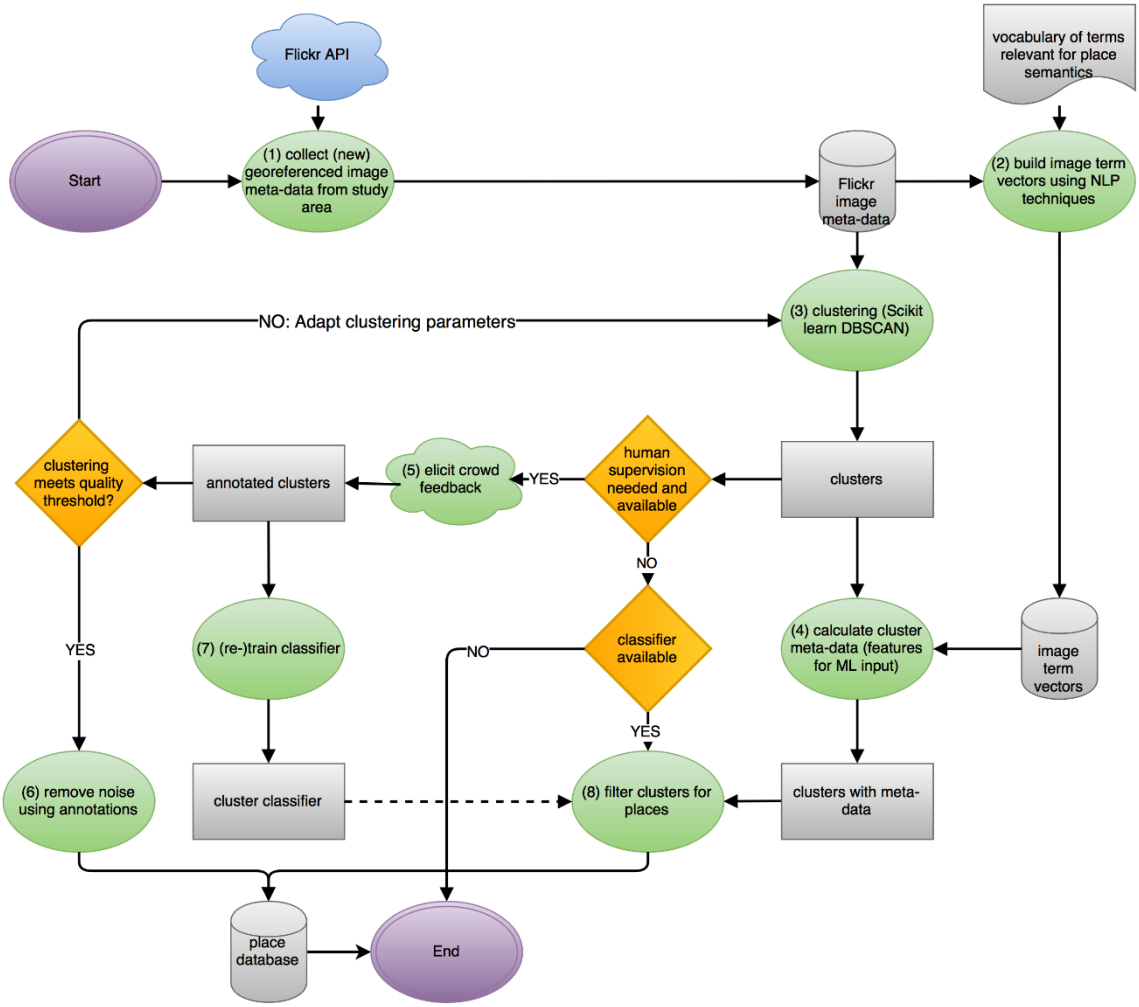

Figure 5 Pilot study prototype implementation and workflow

Phase 3. Find spatial clusters

Images that contain any terms found in our vocabulary are sent to the clusterer module to search for spatial clusters. We implemented the DBSCAN clustering from the Python Scikit-learn framework. DBSCAN can deal well with varying density of points as well as irregular shapes of clusters, and Scikit-learn could be reliably integrated into our workflow. Further, it was computationally inexpensive and fast, and has been used successfully in many studies. The clustering was spatial, i.e. with longitude and latitude as only features.

Phase 4. Compute thematic and spatial cluster characteristics

For each cluster, we calculated several thematic and spatial characteristics, to be used as input features for the ML classifier: First, the average and median cosine similarity. Cosine similarity is a common metric for comparing the semantic similarity between two term vectors, and equals the cosine of the angle between the two vectors. We use it to measure the internal similarity of clusters by calculating the average and median (to mitigate the effect of a single outlier image within an otherwise homogeneous cluster) cosine similarity of all the term vectors for all image pairs. Second, the number of images and unique contributors might indicate which image clusters define a distinct geographic place. Additionally, we computed the average and median silhouette (Rousseeuw, 1987) scores of all items for each cluster. The silhouette coefficient measures how similar an object is compared to the other objects in its cluster. It ranges from -1 to +1, with high values indicating a poor match with other clusters, and a good match with objects in its own cluster.

Phase 5. Crowdsourced labelling and validation of clusters

For the first iteration of the annotation, we consider the found clusters to be potential places based on the single criterion of them being spatially close. The web annotation interface presents these clusters one at a time to a human supervisor. The following screenshot (Figure 6) shows the initial information presented to a human supervisor. The interface then presents a cluster to the supervisor, with the location of the contained images shown on a map (using OpenStreetMap as a base map) located on the left side of the interface, and the actual images shown on the right side (for a screenshot, see Figure 7). It is possible to select images both on the map and gallery. Empty spots in the gallery are images that could not be retrieved, because it had been deleted or the access rights modified since the retrieval of the metadata. Once a supervisor clicks on the “Provide feedback” button, s/he has the opportunity to comment on the spatial layout and thematic consistency of the cluster (Figure 8).

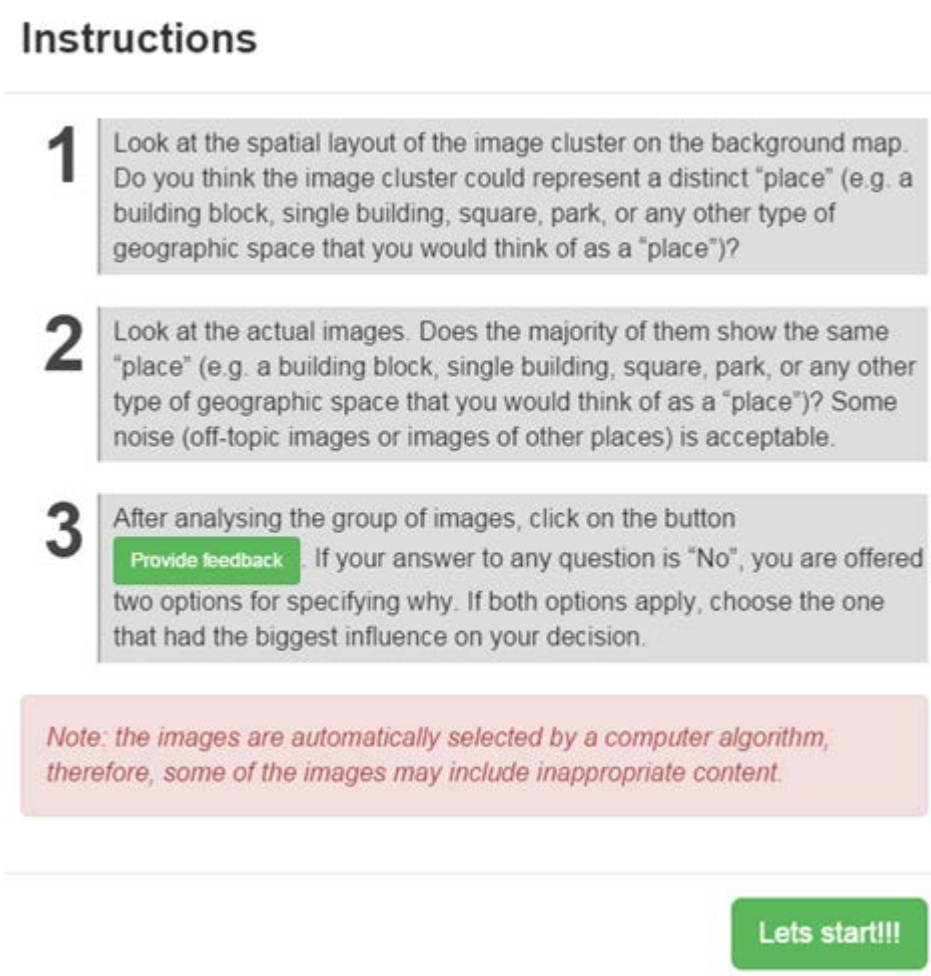

Figure 6 Initial instructions to human supervisors

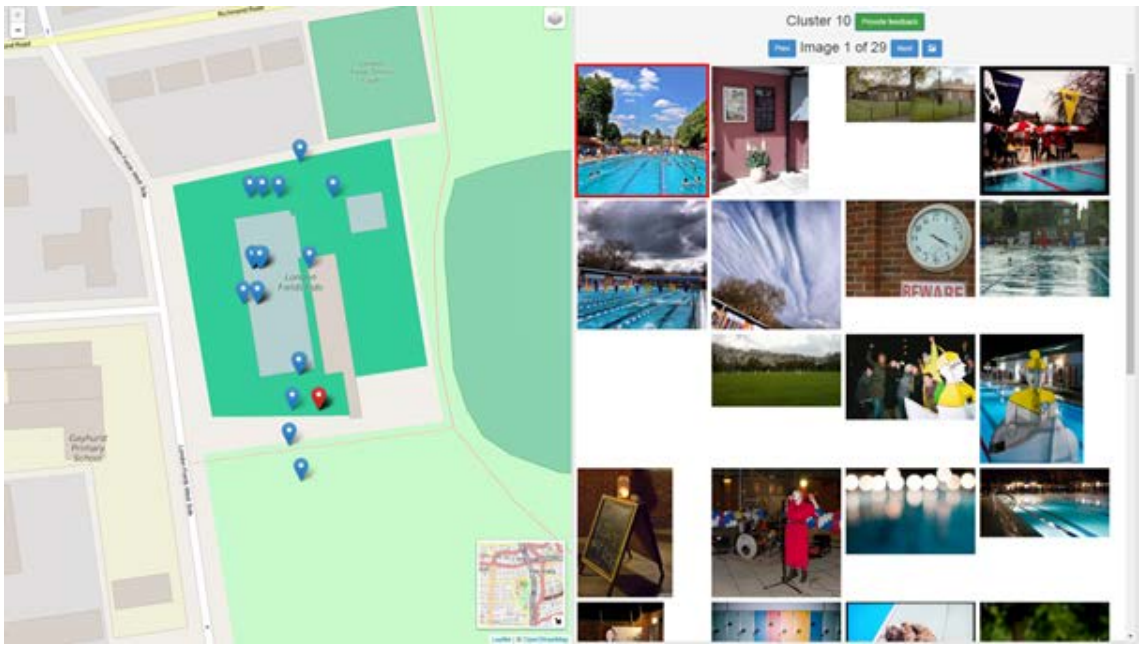

Figure 7 Interface showing typical cluster to be labelled

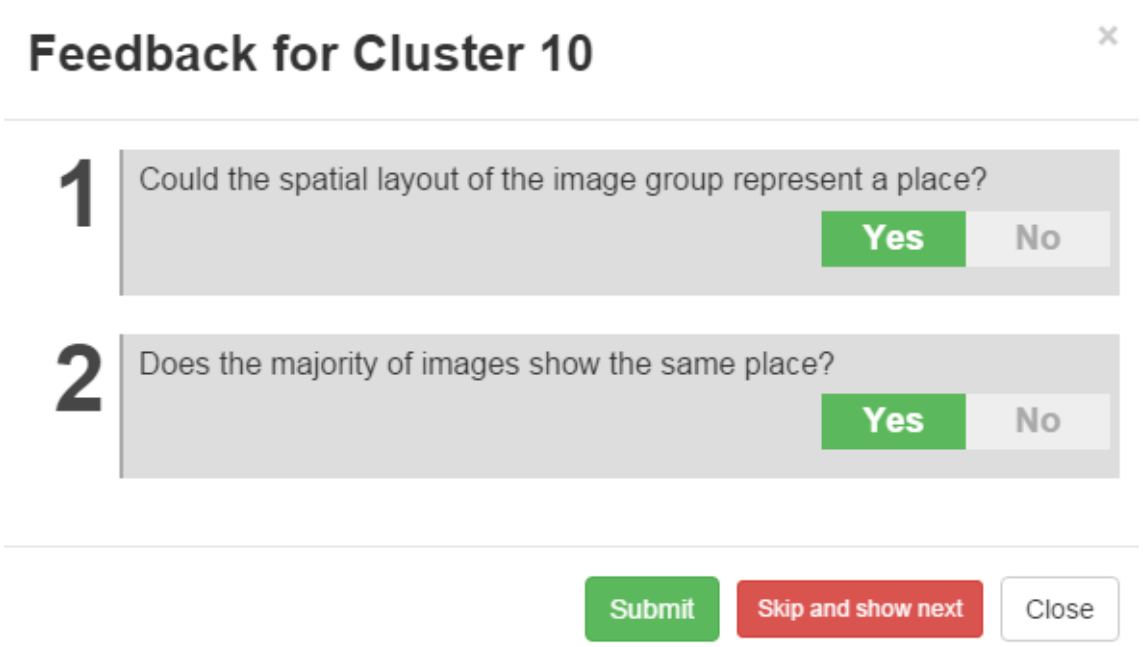

Figure 8 Supervisor feedback for a cluster

If any question is answered "No", additional feedback options appear. For question 1, the supervisor can distinguish between "Wrong shape" and "Too large". An option of "Too small" was discarded after initial tests, since it is impossible to judge whether a cluster is too small and early tests showed that it is also very unlikely to happen. Even if all images are on exactly the same location, their content can cover much more geographic area. For question 2, it is possible to distinguish between "There is more than one place shown" and "There are too many images that are not about a place at all!"

The current interface design relies on past experiments (Ostermann et al., 2013) and a small pre-study, in which users were shown the questions, and then asked what they think that their task would be. Their comments led to a refinement of the introductory text and actual questions, with another round of questioning. The current set-up is based on three such user interviews.

Phase 6: Assess supervision results and remove noise

If the investigator(s) consider the results satisfactory and complete, non-place clusters can be simply removed as noise and the remaining stored in a UGGC place database. If not, the supervisor feedback leads to adjusting the clustering hyper-parameters. For the pilot study, we did not define a stable and satisfactory result. Instead, the first iteration

used hyper-parameters that lead to geographically big clusters, and the second iteration hyper-parameters resulted in smaller (more compact) clusters. The aim was to account for the unknown and varying scale of places, and to test the supervisors' feedback.

Phase 7. (Re-)train a classifier to detect places from noise using the cluster labels and characteristics

The characteristics derived in phase 4 form the feature space for the ML algorithm to classify the clusters into "place" or "non-place". For starters, we choose a simple J48 decision tree learner implemented in Weka[3], which has performed well on previous occasions (Spinsanti and Ostermann, 2013).

Phase 8. Filter clusters using classifier

If there is no or insufficient human supervision available, the system could use a previously trained classifier to filter noise from the clusters. This step is only included for completeness but not implemented in the pilot study.

4.3 Results

The initial dataset retrieved from the Flickr API consists of the meta-data for 5,182,330 geo-referenced photos uploaded up to and including November 2014. If there were no terms in the meta-data matching the vocabulary, the photo was discarded from further analysis, leaving 2,309,760 photos. The GLA contains too many (potential) places for our limited pilot study. Therefore, we extracted the data for seven City wards (Queensbridge, Dalston, Hackney Downs, Leabridge, Victoria, Hackney Central, Chatham), which are not located in the areas most frequented by tourists, and where the administrative boundaries form a mostly convex hull to reduce edge effects for places that lie close to administrative borders. The boundaries of the study area within London are shown in Figure 9. The photo meta-data from this area (n=16632) was then fed into the Scikit DBSCAN clusterer.

[3] http://www.cs.waikato.ac.nz/ml/weka/

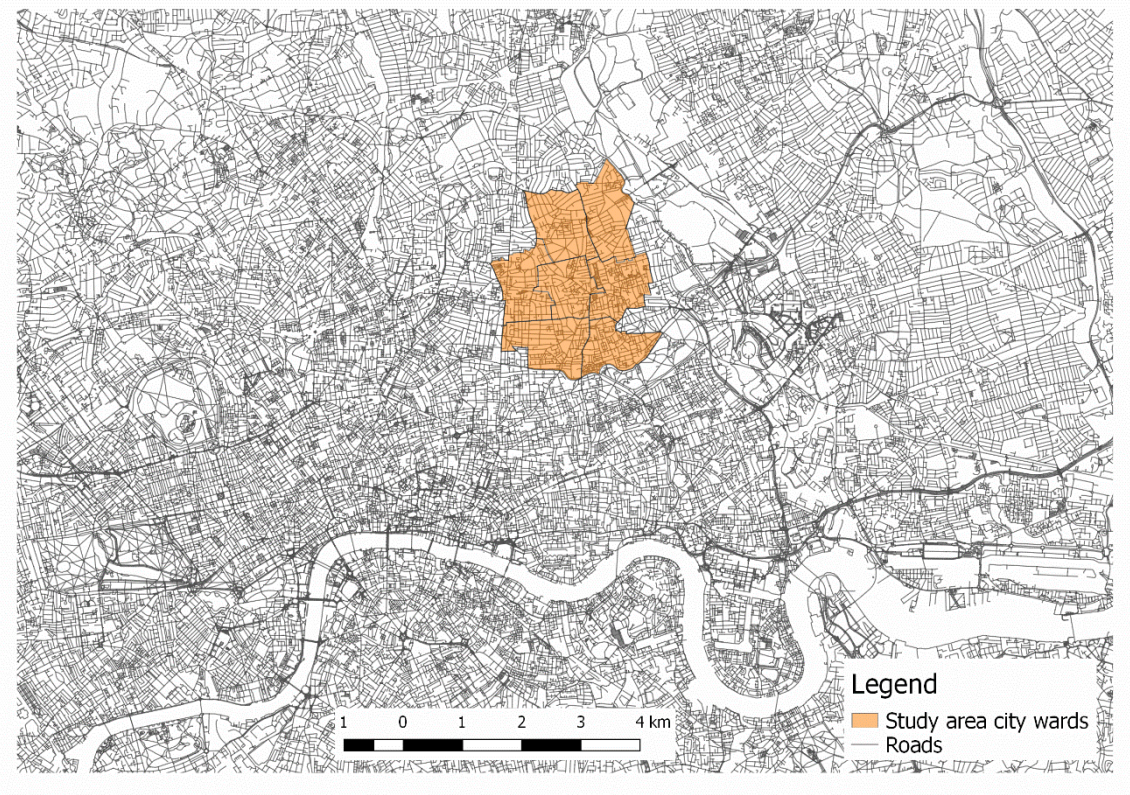

Figure 9 Study area of seven City Wards within London

The initial run used the DBSCAN parameters of eps=0.0005 and a minimum number of 10 images per cluster. The resulting 77 clusters were then shown to the human supervisors (n=5, with some annotators skipping certain clusters where they felt not confident enough to provide feedback) using the web interface described in the previous section. As expected, there is some disagreement between the supervisors. The small number of supervisor allowed a simple majority vote (i.e. the most common answer is assigned to that cluster, with the cluster being dropped from further analysis in case of ties). The results of the first round of annotations are shown in Table 1. These results indicate that 55% of the clusters contain one or more possible places (categories A and C), and 45% contain too much noise.

Table 1 Frequency of annotator responses to first clustering (n = 77 clusters, majority vote in case of inter-rater disagreement, m = 5 annotators; x = 15 excluded if no majority vote available)

Labeling of cluster (Question 1)	Frequency of label	Labeling of place (Question 2)	Frequency of label
0 (spatial cluster correct)	42	A (one place)	22
1 (spatial cluster wrong shape)	14	B (too much off-topic)	28
2 (spatial cluster too big)	6	C (more than one place)	12

In an attempt to filter that noise, the clusters were then used to train a J48 classifier implemented in Weka, using a 10-fold stratified cross-validation (CV) to estimate performance. Removing ambiguous clusters (without majority rater agreement, see above) improves the J48 classification performance significantly. Using all the features described under phase 4, the resulting J48 model correctly classifies on average (of the 10-fold CV) 71% of all instances. The average recall is 79% if we consider only Type II errors (false negatives) to be clusters that contain one or more places – categories A and C – but were classified as noise or category B. The full confusion matrix showing the combined results of the 10-fold CV is shown in Table 2.

Table 2 Confusion matrix of 10-fold CV classification of place-relatedness of clusters

Original Label	Classified as A	Classified as B	Classified as C
Shows one place (A)	17	3	2
Too much off-topic (B)	6	21	1
Shows more places (C)	2	4	6

Many clusters actually consist of more than one place, and some of the clusters are very large in number and spread over a large area (a characteristic of DBSCAN). Therefore in a second iteration, the results were used to modify the clustering hyper-parameters to allow for smaller clusters (DBSCAN parameters of eps = 0.0003 and a minimum of 5 samples per cluster), resulting in a set of 210 clusters as Figure 10 and Table 3 show:

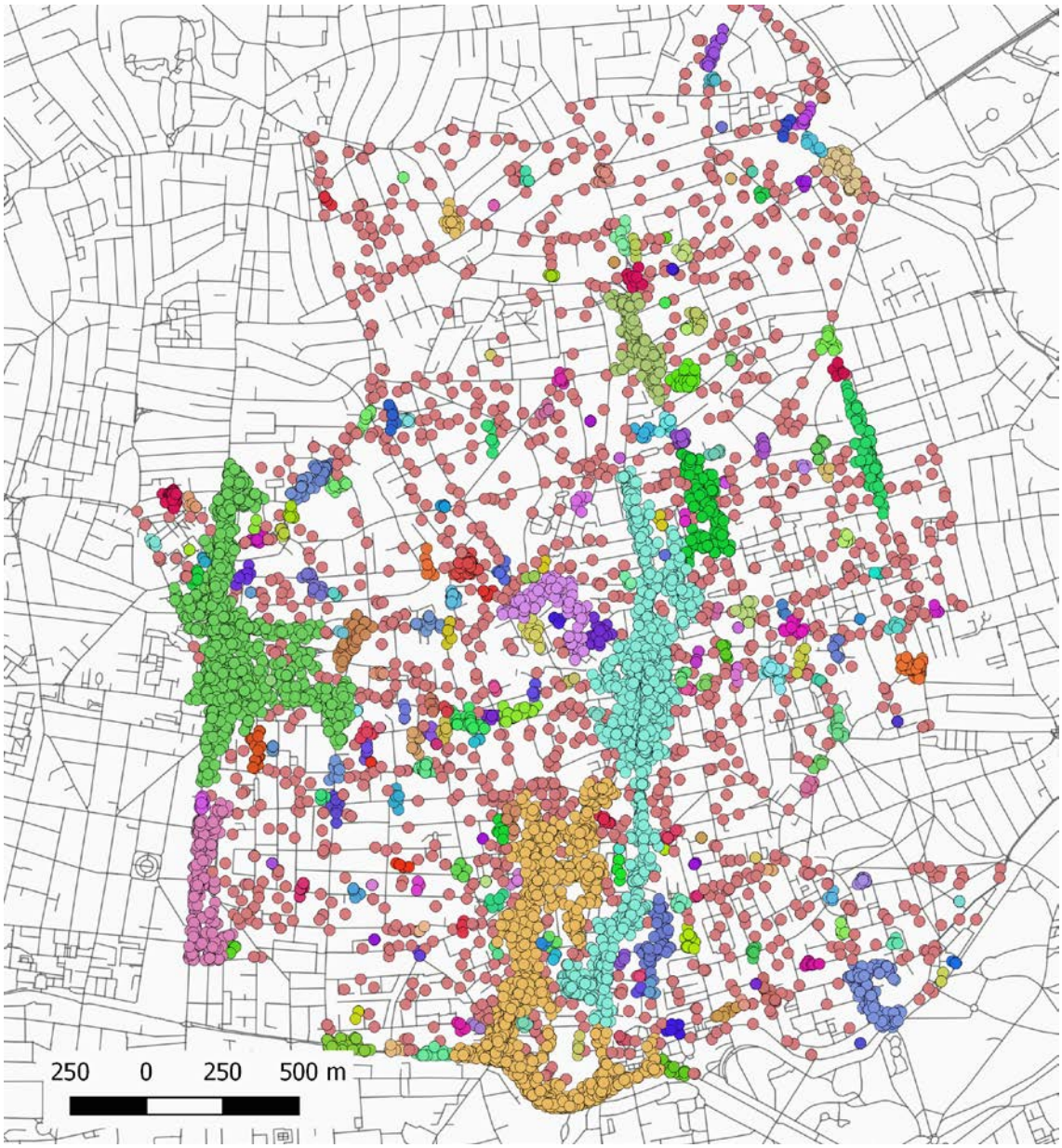

Figure 10 Distribution of images in study area, colored according to cluster attribution

Table 3 Frequency of annotator responses to second clustering (n=210 clusters, one annotator)

Labeling of cluster (Question 1)	Frequency of label	Labeling of place (Question 2)	Frequency of label
0 (spatial cluster correct)	189	A (one place)	91
1 (spatial cluster wrong shape)	11	B (too much off-topic)	101
2 (spatial cluster too big)	10	C (more than one place)	18

The ratio of signal-to-noise remains similar: 52% of the clusters contain one or more places, and 48% contain too much noise or no places at all. Applying the original classifier to predict the second set results in an overall decrease of classification performance: only 49% are correctly classified, with a recall of only 45%, indicating a large number of false negatives. Some degradation is to be expected, given that the original training data set was created with different clustering hyper-parameters, but the number of Type II errors is not acceptable. We cannot use the original classifier to predict on new data generated with the changed spatial clustering hyper-parameters. The full confusion matrix is shown in Table 4.

Table 4 Confusion matrix for prediction of place-relatedness of clusters, first classifier, second iteration

Original Label	Classified as A	Classified as B	Classified as C
Shows one place (A)	38	45	8
Too much off-topic (B)	28	63	10
Shows more places (C)	1	15	2

However, training a new J48 model using the second dataset also results in much lower performance than the first one: The performance estimation of stratified 10-fold CV is only 49.5% correctly classified instances, with a recall (again categories A and C combined) of 66%. The full confusion matrix is shown in Table 5:

Table 5 Confusion matrix of 10-fold CV classification of place-relatedness of clusters, second classifier, second iteration

Original Label	Classified as A	Classified as B	Classified as C
Shows one place (A)	58	29	4
Too much off-topic (B)	53	44	4
Shows more places (C)	8	8	2

As a result, classification performance has overall degraded with the spatial clustering at a finer granularity. Although much of this is due to misclassified off-topic clusters (false positives), the number of false negatives (misclassified as non-place related) is too high. An attempt to address this issue by combining the two cluster sets, adding the cluster hyper-parameters as classification features, and training a new classifier did not improve the expected error.

5 Discussion

The implementation of this workflow in a pilot study produced solid results with established and well-known analysis methods. An initial, purely unsupervised DM approach to detect places produced too many cluster candidates (potential places), emphasizing the need for reduction to meaningful places through ML. The large number of instances to label, and the iterative character of the search for good hyper-parameters support our approach of crowdsourced supervision of training/learning. Because of pilot character of the study, we limit it to a manageable number of images (and resulting clusters), but in principle this scales well computationally and organizationally. The computational costs were sufficiently low so that all scripts would run sufficiently fast (less than a minute of run-time) on a common-off-the-shelf business laptop. The software used in the process (PostGIS database, Python scripting and Weka) are mature enough so that a user with a moderate amount of IT-skills can set it up within few hours. All these criteria make the approach suitable for citizen science projects not having a strong or dedicated computer or data science expertise. The web interface proved easy enough to navigate and work with. Annotator feedback indicated that the questions could be formulated clearer. This

is a result of the chosen systems-centered design perspective. A user-centered design process (Yovcheva et al., 2013) was not possible due to temporal and resource constraints, but remains an objective for future iterations. The same holds true for the small number of supervisors. However, we consider both limitations acceptable for the initial pilot study, since we do not yet plan to implement advanced decision-making or analysis functionality. Human labelling required less than a minute per cluster (although we did not account for varying screen sizes - the bigger the screen, the less scrolling to depict all images from large clusters, hence a possible reduction in annotation time). The initial clustering was good enough to result in good inter-rater agreement.

Regarding first criterion for a hybrid information processing workflow (see section 3.1), the system architecture does not yet allow true stream processing. While micro-batching can be implemented through scheduled tasks, near real-time processing will require changes to the architecture. Currently, all processes are implemented as chained Python scripts running on a laptop. For future implementations, stream processing frameworks such as Apache Storm or Spark, running on a cloud platform, promise the biggest flexibility and reliability. This would also reduce the amount of manual work needed, and increase the stability of the system. The clustering will need to be spatially bounded (i.e. a single new UGGC should not trigger a re-clustering of the whole study area). The ML part could profit from active learning strategies. In active learning, a learner chooses instances to be labelled and presents them to the human annotator, with the aims of maximizing the impact of human annotation and remaining flexible towards new instances. However, there is also evidence that passive learning is better suited for annotation by domain novices, while active learning profits in particular from domain expertise, and batch-mode labeling is better suited to multiple, parallel annotators (Settles, 2009). The crowdsourced supervision and constant updating of training sets for the learning algorithms is possible using the Pybossa framework. Additionally, the current convenience sample of recruiting annotators has to be replaced with a more systematic and sustainable process. To do so, we can rely on current research on establishing successful and lasting collaborative frameworks (Eveleigh et al., 2014).

Regarding the other criteria (see 3.1), the modular structure allows to plug-in different data sources when needed, although this is currently still a manual process. Adaptations for geosocial media platforms such as Twitter that provide a streaming API will

require an additional layer as mentioned in the previous section. However, a fully web-based analysis engine relying on web resources and WPS is possible. All analysis has been conducted with open source software, openly available algorithms, and it is freely available as supplementary material.

While the pilot study fulfilled well its primary aim of demonstrating a feasible approach to hybrid geoinformation, its secondary aim of searching for meaningful places suffered from the clustering and learning performance, which leave substantial room for improvement, both for spatial and thematic dimensions. A considerable share of images have not been assigned a cluster or a mega-clusters covering large parts of the study area. Taking the temporal dimension into account (Birant and Kut, 2007) might help to detect ephemeral events and distinguish them from persistent features. Although one would expect for a valid thematic cluster a high average cosine similarity and several contributors, there are some outliers that make this simple classification very difficult. Further, finding places through UGGC is a complex task, probably requiring more features. Tests with more features using individual terms aggregated per cluster and different classifiers (Adaboost M1, Naive Bayes) show no improvement. It is likely that ancillary data from other UGGC sources or socio-demographic data from authoritative sources could help. This also leads to questioning the choice of input data and study objectives, because many images have only very few and quite generic terms in their textual descriptions, making the extraction of place characteristics purely from UGGC very difficult. Another UGGC source or target objective could make the actual analysis easier and help to focus on the methodological aspects of cognitive task and workflow modeling.

Another future area of investigation is the suitability of our approach to reduce an analytic divide between those who contribute information and those who process it. We argue that this analytical divide can impact negatively on the meaningfulness of the analysis results, the potential for empowerment of marginal groups, and even delegitimize democratic processes (Helbing and Pournaras, 2015). Further, the increasing reliance of science on massive amounts of data, also described as the 4th scientific paradigm of data-intensive science (Hey et al., 2009), should not lead to a return of positivism. A higher level of participation of citizens in the analytical process can lead to improved autonomy, new knowledge and higher societal relevance of research output (Feyerabend, 1993).

6 Conclusions and Outlook

In this paper, we have addressed the question on how to include crowdsourced supervision into an analytical workflow to mitigate some of the challenges associated with the processing of UGGC. We first evaluated the challenges of processing UGGC, followed by an evaluation of the advantages and problems of two basic approaches, i.e. human curation and machine computing. We have shown that these have complementary strengths and weaknesses, and combining the two promises to improve processing of UGGC. Further, we have developed a model workflow of hybrid geoinformation processing for crowdsourcing the supervision of geospatial analysis tasks. This hybrid geoinformation processing workflow integrates crowdsourced supervision in an iterative manner. It distinguishes between phases that profit from a geospatial analysis (georeferencing, clustering), and those that can function without explicitly geospatial components (content classification).

We have then defined suitability criteria for analysis tasks and techniques to elicit knowledge from supervisor and feeding it back into the system, formalized the tasks in a bottom-up basic classification for future re-use, and mapped concrete geospatial analysis tasks to querying human supervisor: The hyper-parameterization of a DM (clustering) task and of a ML (classification) task.

Finally, we have implemented a prototype in a pilot study to demonstrate the feasibility of the approach. It relies exclusively on established and available open source software and algorithms, and implements major parts of the proposed workflow model. First, it collects and stores UGGC from a photo-sharing platform. It then enriches it with additional information (term vectors), before it clusters them using the DBSCAN algorithm. The resulting clusters are presented in a web-interface that allows asynchronous annotation by multiple human supervisors. The responses are stored and can be used to improve the clustering by adjusting the hyper-parameters or choosing a different clustering algorithm, and to classify resulting clusters into those about a place or not.

The pilot study highlighted several issues that future research should address: The mapping of tasks to supervision needs further formalization and expansion, e.g. by graph representation such as the semantic wiki version of the GIS&T BoK (BokWIKIEx-

plorer[4]). Additionally, UGGC requires a system architecture that supports stream processing. Finally, the crowdsourced supervision needs a sustainable organization, so that more training sets can be labeled. Particularly promising approaches are active learning and online learning. The developed conceptual model workflow, task classification, and prototype are first steps towards the larger goal of crowdsourced analysis and supervision. The model will profit from further refinement and expansion, while the prototype's components are intentionally kept simple and offer much room for improvement and optimization. A possible future case studies is the collaborative supervision of classification of UGGC into credible or untrustworthy for the critical task of assigning resources in disaster response.

Acknowledgements

We thank the annotators and interface testers for their efforts and valuable feedback.

7 References

Abel, F., Hauff, C., Houben, G.-J., Stronkman, R., Tao, K., 2012. Twitcident: Fighting Fire with Information from Social Web Stream, in: International Conference on Hypertext and Social Media, Milwaukee, USA. ACM.

Barrington, L., Turnbull, D., Lanckriet, G., 2012. Game-powered machine learning. Proc. Natl. Acad. Sci. U. S. A. 109, 6411–6416.

Birant, D., Kut, A., 2007. ST-DBSCAN: An algorithm for clustering spatial–temporal data. Data Knowl. Eng. 60, 208–221. doi:10.1016/j.datak.2006.01.013

Bonney, R., Cooper, C.B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K.V., Shirk, J., 2009. Citizen Science: A Developing Tool for Expanding Science Knowledge and Scientific Literacy. BioScience 59, 977–984. doi:10.1525/bio.2009.59.11.9

Burger, J.D., Henderson, J., Kim, G., Zarrella, G., 2011. Discriminating gender on Twitter, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Edinburgh, United Kingdom, pp. 1301–1309.

Butler, D., 2013. When Google got flu wrong. Nature 494, 155–156.

[4] http://gistbok.org/bokwiki/index.html

Camponovo, M.E., Freundschuh, S.M., 2014. Assessing uncertainty in VGI for emergency response. Cartogr. Geogr. Inf. Sci. 41, 440–455. doi:10.1080/15230406.2014.950332

Caquard, S., 2014. Cartography II: Collective cartographies in the social media era. Prog. Hum. Geogr. 38, 141–150. doi:10.1177/0309132513514005

Craglia, M., Ostermann, F.O., Spinsanti, L., 2012. Digital Earth from vision to practice: making sense of citizen-generated content. Int. J. Digit. Earth 5, 398–416. doi:10.1080/17538947.2012.712273

D'Hondt, E., Stevens, M., Jacobs, A., 2013. Participatory noise mapping works! An evaluation of participatory sensing as an alternative to standard techniques for environmental monitoring. Pervasive Mob. Comput. 9, 681–694. doi:10.1016/j.pmcj.2012.09.002

Dodge, M., Kitchin, R., 2013. Crowdsourced cartography: mapping experience and knowledge. Environ. Plan. A 45, 19–36.

Eveleigh, A., Jennett, C., Blandford, A., Brohan, P., Cox, A.L., 2014. Designing for Dabblers and Deterring Drop-outs in Citizen Science, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '14. ACM, New York, NY, USA, pp. 2985–2994. doi:10.1145/2556288.2557262

Fast, V., Rinner, C., 2014. A Systems Perspective on Volunteered Geographic Information. ISPRS Int. J. Geo-Inf. 3, 1278–1292.

Feick, R., Robertson, C., 2014. A multi-scale approach to exploring urban places in geotagged photographs. Comput. Environ. Urban Syst. doi:10.1016/j.compenvurbsys.2013.11.006

Feyerabend, P., 1993. Against method. Verso.

Gahegan, M., 2003. Is inductive machine learning just another wild goose (or might it lay the golden egg)? Int. J. Geogr. Inf. Sci. 17, 69–92. doi:10.1080/713811742

Garcia-Martí, I., Zurita-Milla, R., Swart, A., van den Wijngaard, K.C., van Vliet, A.J.H., Bennema, S., Harms, M., 2016. Identifying Environmental and Human Factors Associated With Tick Bites using Volunteered Reports and Frequent Pattern Mining: Environmental and Human Factors Associated with Tick Bites. Trans. GIS. doi:10.1111/tgis.12211

Geographic Information Science and Technology Body of Knowledge, 2006. . Association of American Geographers, Washington, D.C.

Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography. GeoJournal 69, 211–221.

Graham, M., Shelton, T., 2013. Geography and the Future of Big Data, Big Data and the Future of Geography. Dialogues Hum. Geogr. 3, 255–261.

Granell, C., Ostermann, F.O., 2016. Beyond data collection: Objectives and methods of research using VGI and geo-social media for disaster management. Comput. Environ. Urban Syst. doi:10.1016/j.compenvurbsys.2016.01.006

Harvey, F., 2013. To Volunteer or to Contribute Locational Information? Towards Truth in Labeling for Crowdsourced Geographic Information, in: Sui, D., Elwood, S., Goodchild, M. (Eds.), Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice. Springer, Berlin, pp. 31–42.

Haworth, B., 2016. Emergency management perspectives on volunteered geographic information: Opportunities, challenges and change. Comput. Environ. Urban Syst. 57, 189–198. doi:10.1016/j.compenvurbsys.2016.02.009

Helbing, D., Pournaras, E., 2015. Society: Build digital democracy. Nature 527, 33–34. doi:10.1038/527033a

Hey, A., Tansley, S., Tolle, K., 2009. The Fourth Paradigm: Data-Intensive Scientific Discovery, The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research.

Hollenstein, L., Purves, R.S., 2010. Exploring place through user-generated content: Using Flickr tags to describe city cores. J. Spat. Inf. Sci. 21.

Imran, M., Castillo, C., Lucas, J., Meier, P., Rogstadius, J., 2014. Coordinating Human and Machine Intelligence to Classify Microblog Communications in Crises, in: Proceedings of the 11th International ISCRAM Conference. Presented at the 11th International ISCRAM Conference, ISCRAM.

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., Meier, P., 2013. Practical Extraction of Disaster-Relevant Information from Social Media, in: Car, L., Laender, A.H.F., Loscio, B.F., King, I., Fontoura, M., Vrandeeiae, D., Oliveira, J.P.M. de, Lima, F., Wilde, E. (Eds.), WWW2013 Companion Publication. Presented at the 22nd International World Wide Web Conference, IW3C2, Rio de Janeiro, Brazil, p. 4.

Kaiser, C., Pozdnoukhov, A., 2013. Enabling real-time city sensing with kernel stream oracles and MapReduce. Spec. Issue Pervasive Urban Appl. 9, 708–721. doi:10.1016/j.pmcj.2012.11.003

Kamar, E., Hacker, S., Horvitz, E., 2012. Combining human and machine intelligence in large-scale crowdsourcing, in: Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 1. International Foundation for Autonomous Agents and Multiagent Systems, Valencia, Spain, pp. 467–474.

Kanevski, M., Pozdnoukhov, A., Timonin, V., 2009. Machine learning for spatial environmental data: theory, applications and software. EPFL Press ; Distributed by CRC Press, Lausanne, Switzerland; Boca Raton, Fla.

Kanevski, M., Pozdnoukhov, A., Timonin, V., 2008. Machine Learning Algorithms for GeoSpatial Data. Applications and Software Tools, in: Proceedings of the 4th Biennial Meeting of iEMSs. Presented at the International Congress on Environmental Modelling and Software Integrating Sciences and Information Technology for Environmental Assessment and Decision Making.

Kelley, M.J., 2013. The emergent urban imaginaries of geosocial media. GeoJournal 78, 181–203. doi:10.1007/s10708-011-9439-1

Kelso, N.V., 2010. Haiti OpenStreetMaps + Google Map Maker. Kelsos Corner.

Kessler, C., Hendrix, C., 2015. The Humanitarian eXchange Language: Coordinating disaster response with semantic web technologies. Semantic Web 6, 5–21. doi:10.3233/SW-130130

Kuhn, W., 2012. Core concepts of spatial information for transdisciplinary research. Int. J. Geogr. Inf. Sci. 26, 2267–2276. doi:10.1080/13658816.2012.722637

Kullenberg, C., Kasperowski, D., 2016. What Is Citizen Science? – A Scientometric Meta-Analysis. PLOS ONE 11, e0147152. doi:10.1371/journal.pone.0147152

Liu, S.B., 2014. Crisis Crowdsourcing Framework: Designing Strategic Configurations of Crowdsourcing for the Emergency Management Domain. Comput Support. Coop Work 23, 389–443.

Meier, P., 2013. Humanitarianism in the Network Age: Groundbreaking Study | iRevolution. iRevolution.

Morrow, N., Mock, N., Papendieck, A., Kocmich, N., 2011. Independent Evaluation of the Ushahidi Haiti Project. DISI - Development Information Systems International.

Ostermann, F.O., Huang, H., Andrienko, G., Andrienko, N., Capineri, C., Farkas, K., 2015. Extracting and comparing places using geo - social media, in: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci., II-3/W5. Presented at the ISPRS Geospatial Week 2015, ISPRS, La Grande Motte, France, pp. 311–316. doi:10.5194/isprsannals-II-3-W5-311-2015

Ostermann, F.O., Spinsanti, L., 2012. Context Analysis of Volunteered Geographic Information from Social Media Networks to Support Disaster Management: A Case Study On Forest Fires. Int. J. Inf. Syst. Crisis Response Manag. 4, 16–37. doi:10.4018/jiscrm.2012100102

Ostermann, F.O., Tomko, M., Purves, R., 2013. User Evaluation of Automatically Generated Keywords and Toponyms for Geo-Referenced Images. J. Am. Soc. Inf. Sci. Technol. 64, 480–499. doi:10.1002/asi.22738

Purves, R., Edwardes, A., Wood, J., 2011. Describing place through user generated content. First Monday Vol. 16 Number 9 - 5 Sept. 2011.

Rattenbury, T., Naaman, M., 2009. Methods for extracting place semantics from Flickr tags. ACM Trans. Web 3.

Rousseeuw, P.J., 1987. Silhouettes - a Graphical Aid to the Interpretation and Validation of Cluster-Analysis. J. Comput. Appl. Math. 20, 53–65.

Santos, J., Anastácio, I., Martins, B., 2015. Using machine learning methods for disambiguating place references in textual documents. GeoJournal 80, 375–392. doi:10.1007/s10708-014-9553-y

Settles, B., 2009. Active Learning Literature Survey (Computer Sciences Technical Report No. 1648). University of Wisconsin, Madison.

Sigurbjörnsson, B., Van Zwol, R., 2008. Flickr tag recommendation based on collective knowledge, in: Proceedings of the 17th International Conference on World Wide Web. Presented at the 17th International Conference on World Wide Web, ACM Press, Beijing, China, pp. 327–336.

Spinsanti, L., Ostermann, F.O., 2013. Automated geographic context analysis for volunteered information. Appl. Geogr. 43, 36–44. doi:10.1016/j.apgeog.2013.05.005

Stefanidis, A., Crooks, A., Radzikowski, J., 2013. Harvesting ambient geospatial information from social media feeds. GeoJournal 78, 319–338. doi:10.1007/s10708-011-9438-2

Sui, D., Elwood, S., Goodchild, M.F. (Eds.), 2012. Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice. Springer, Berlin.

Sun, C., Rampalli, N., Yang, F., Doan, A., 2014. Chimera: large-scale classification using machine learning, rules, and crowdsourcing. Proc. VLDB Endow. 7, 1529–1540. doi:10.14778/2733004.2733024

Thrift, N., French, S., 2002. The automatic production of space. Trans. Inst. Br. Geogr. 27, 309–335. doi:10.1111/1475-5661.00057

Tversky, B., Hemenway, K., 1983. Categories of environmental scenes. Cognit. Psychol. 15, 121–149.

Van Canneyt, S., Schockaert, S., Van Laere, O., Dhoedt, B., 2012. Detecting Places of Interest Using Social Media. IEEE, pp. 447–451. doi:10.1109/WI-IAT.2012.19

Yovcheva, Z., van Elzakker, C.P.J.M., Köbben, B., 2013. User requirements for geo-collaborative work with spatio-temporal data in a web-based virtual globe environment. Appl. Ergon. 44, 929–939. doi:10.1016/j.apergo.2012.10.015