Service-based combination of quality assurance and fusion processes for the validation of crowdsourced observations

Stefan Wiemann
Technische Universität Dresden
Geoinformation Systems
Dresden, Germany
stefan.wiemann@tu-dresden.de

Sam Meek
University of Nottingham
Geospatial Institute
Nottingham, UK
sam.meek@nottingham.ac.uk

Colin Chapman
Welsh Government
Land, Nature and Forestry Division
Aberystwyth, UK
colin.chapman@wales.gsi.gov.uk

Didier G Leibovici
University of Nottingham
Geospatial Institute
Nottingham, UK
didier.leibovici@nottingham.ac.uk

Jamie Williams
Environment Systems

Aberystwyth, UK
jamie.williams@envsys.co.uk

Mike J Jackson
University of Nottingham
Geospatial Institute
Nottingham, UK
mike.jackson@nottingham.ac.uk

Lars Bernard
Technische Universität Dresden
Geoinformation Systems
Dresden, Germany
lars.bernard@tu-dresden.de

Abstract

This paper deals with the combination of web based quality assurance and data fusion processes to support mutual validation of crowdsourced observations and authoritative data. Results stem from the EU FP7 funded Citizen Observatory Web (COBWEB) project. COBWEB targets new methods and tools for collecting Citizen Science and Crowdsourced data for policymaking. Major issues relate to uncertainty in the quality and accuracy of such data. We present a design for a framework and an implementation that incorporates techniques to qualify and add value to crowdsourced data with a view to addressing some of the uncertainty and isolation issues with these sources. We show an example of a Citizen Observatory use case that utilizes data on the distribution of Japanese Knotweed collected by volunteers via a mobile app. Following a service-oriented design, the approach offers generic components which could also be transferred to similar applications for spatial data validation.

Keywords: crowdsourcing, quality assurance, data fusion, ground-truthing.

1 Introduction

The Citizen Observatory Web (COBWEB) project, an FP7 EU funded project, aims at collecting Citizen Science and crowdsourced data for policy making. A major concern for policy makers is assessing the applicability of crowdsourced data as a suitable form of evidence to support decision making or policy formation. Methods to identify and at least document the uncertainty concerning data quality and data accuracy are key to support the usage of these data sources.

This paper builds upon last year's AGILE publications by [11] and [19] to describe a flexible and reusable approach for service-based data quality assurance (QA) and spatial data fusion for the validation of crowdsourced data.

A COBWEB co-design use case, namely the detection of Japanese Knotweed, hereinafter referred to as knotweed, in the Snowdonia National Park in Wales, is taken to demonstrate the feasibility of the approach. Knotweed is an invasive species and known for interfering with native wildlife and causing damage to buildings, roads and water infrastructure [3]. The government thus requires that it is properly removed when it has been identified. The use case includes both the validation of crowdsourced observations and the enhancement of an official knotweed distribution map with validated observations. The aim is to reduce uncertainty of the observations and the distribution map as much as possible to finally create a reliable knotweed distribution map for further analysis and to support decision making.

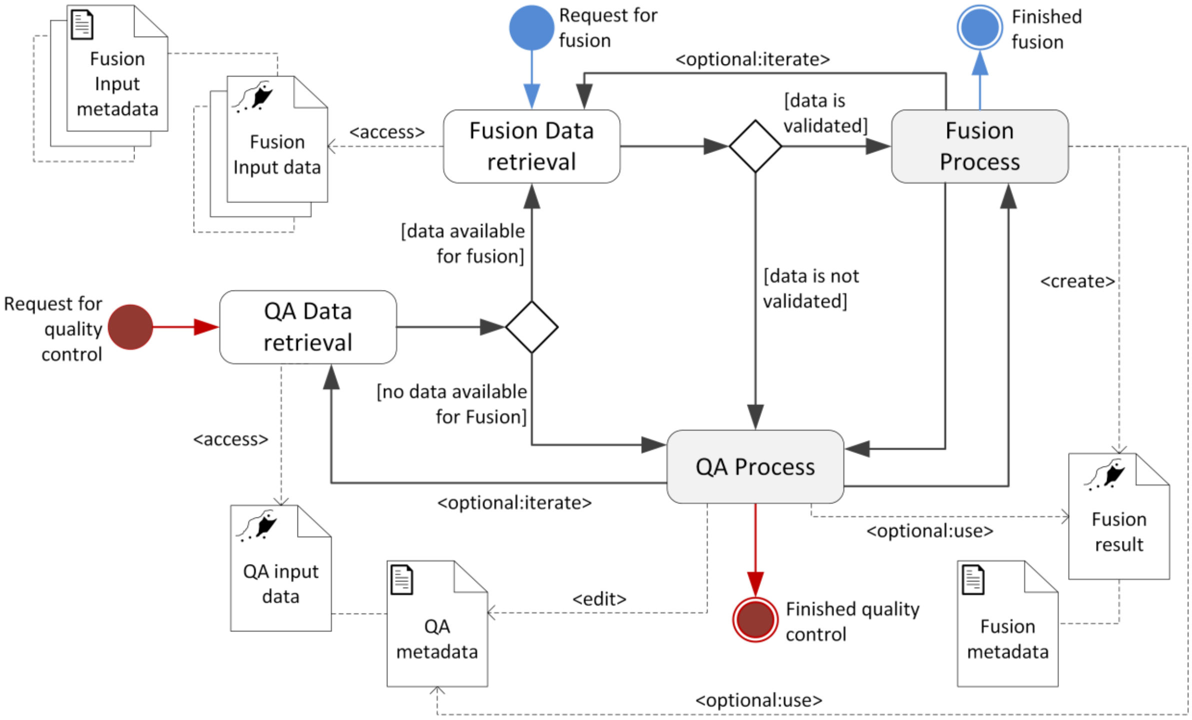
The article introduces application patterns for the combination of QA and data fusion processes in the context of validating of crowdsourced observations (section 2). The application use case is described (section 3) together with a prototype implementation (section 4) to illustrate the application of the presented approach. A final conclusion discusses the current status and identifies future research needs (section 5).

2 Service-based quality assurance and fusion

Data collected by the crowd through formal or informal methods often has an unknown quality and the perception is that it is not suited for use in scientific experimentation or policymaking. Various approaches have been considered for QA of crowdsourced data. Examples include the use of experts [1, 14], the use of a majority decision [8], ontology based methods [18], reliance on multiple observations [6] or the use of polarity and lexical classifiers [9].

In the context of crowdsourced observations, quality assurance is required to assess the quality and reliability of an observation. In terms of quality measures, internal quality and external quality measures are distinguished. Internal quality procedures measure whether a piece of data conforms to a set

Figure 1: Interaction pattern between QA and data fusion processes.



of metrics such as accuracy, precision and consistency with respect to a particular standard. External quality includes measures that describe whether a piece of data is fit for purpose and refers to the comparison with other observations and external data sources. Many of these measures are recorded in standard metadata profiles such as ISO19157 [10] and include required entries *Spatial Accuracy* and *Completeness.* Measures such as *Completeness* may be required as a method of assessing internal quality, as assessment of whether a dataset has all expected entries is useful for data collection agencies. However, a dataset may not necessarily need to be complete to be considered *fit for purpose* depending on the use case.

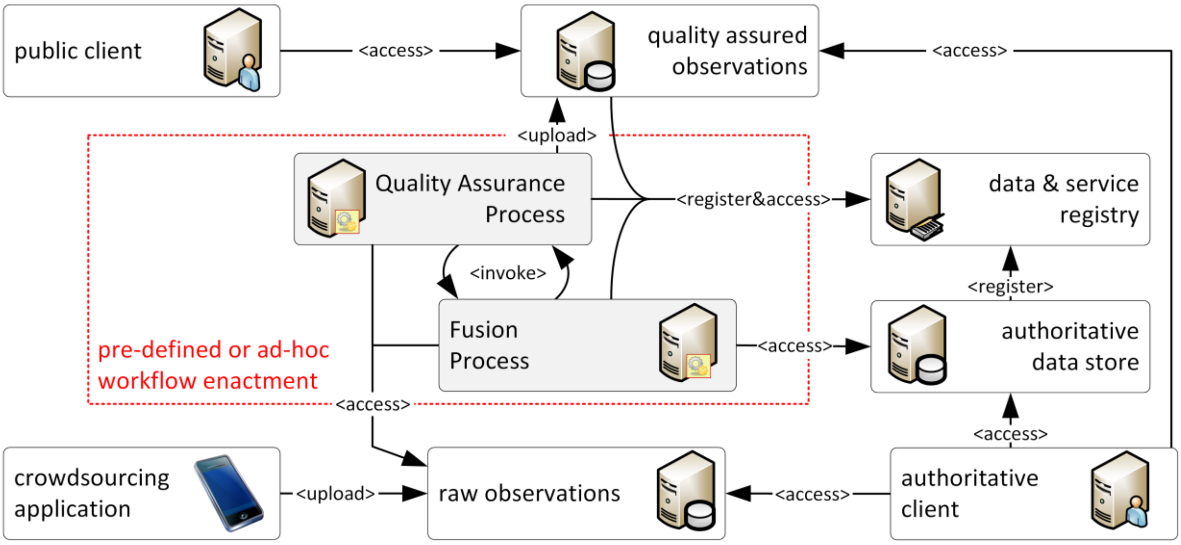
Quality assuring crowd data has largely taken place within specific domains. [11] propose a generic solution and organise quality control tests and checks into *Seven Pillars*, each representing a superclass of processes that are able to assess a specific aspect of the quality of submitted data. Additionally [12] describe a service-based architecture for quality assuring crowd data that utilizes a workflow engine [2] to form a QA procedure for a specific use case from a set of generic QA processes in a service composition. The prototype is based on current OGC (Open Geospatial Consortium) standards and as such algorithms exposed via the Web Processing Service and data made available through OGC data download services. This approach enables maximum flexibility and interoperability and thus allows addressing different use cases for crowdsourcing and citizen science projects.

Besides quality assessment and assuring, data fusion plays an important role for the combination of distributed data sources. Spatial data fusion refers to both, the conflation of homologous objects and the enrichment of data with contextual information. In the geospatial domain, a variety of approaches on spatial data fusion exist, addressing different aspects such as feature similarity measurements [17, 22], dataset matching [13, 15, 21] or linked geospatial data [16]. [19] propose a service-based implementation utilizing OGC Web Services combined with Linked Data concepts.

Figure 1 shows the synergies and mutual benefits being achieved from the combination of QA and data fusion processes. One the one hand, QA can be used to investigate data quality and thereby enhance the fusion process and results. Information on the quality of input data enables uncertainty management and a more reliable accuracy assessment during the fusion process. Derived lineage information significantly increases the applicability of fusion results for further processing. On the other hand, fusion processes can be used to link data sources for external validation purposes in QA. Such a comparison of different datasets has frequently been applied for the validation of crowdsourced data [5, 7]. Thus, at all levels linking to existing data sources is most valuable for the application and further use of the crowdsourced spatial data and should be established wherever feasible.

The architecture and components for distributed QA and data fusion of crowdsourced observations is shown in Figure 2. The main workflow yields quality assured observations that provide reliable information on the observed phenomenon. For recurring processes, two workflow enactment strategies are possible: 1) running a pre-defined QA and fusion process, specifically set up for a specific application or survey, or 2) running a generic, application agnostic workflow in an ad hoc manner. Both require well-defined interfaces, sufficient process descriptions and a process modelling engine to achieve automation.

Figure 2: Components for the validation and fusion of crowdsourced observations.



3 The citizen observatory use case

An application use case has been developed to demonstrate and illustrate the benefits for the combination of service-based QA and data fusion processes. It is aligned to a COBWEB co-design projects led by Welsh Government and deals with the distribution of Japanese Knotweed (Fallopia Japnonica, Figure 3) in the Snowdonia National Park (Wales, UK). The estimated costs for managing the impact of knotweed in Wales are £8.8m a year [20]. Therefore, the UK Environment Agency identifies a number of actions that have to be taken to manage and reduce spread in early stages [4].

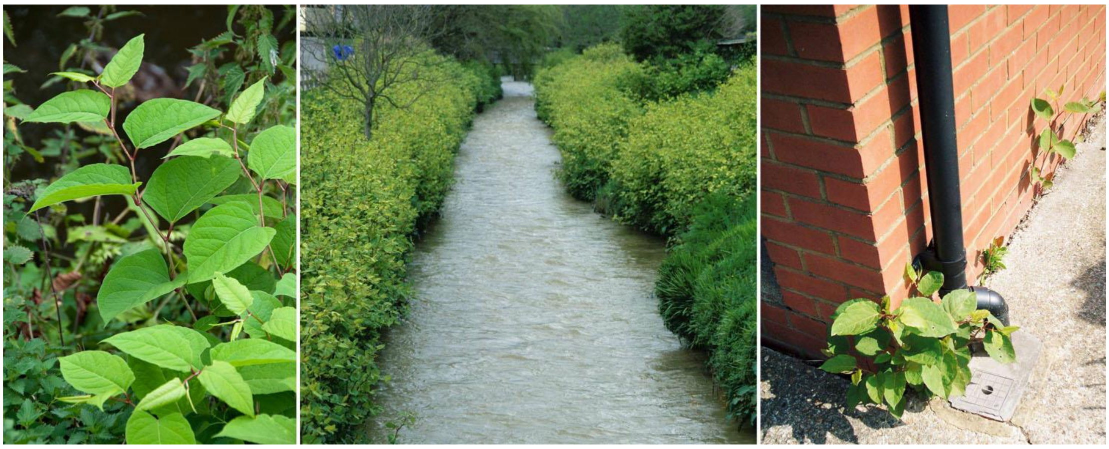
Currently available distribution maps as provided by the National Biodiversity Network (NBN) show the occurrence of knotweed on a 1km resolution (Figure 4). Although, this map will be geometrically and thematically enhanced by a more detailed classification based on color infra-red imagery and LiDAR (Light detection and ranging) in the near future (methodology described in [3]), ground truth observations will remain an essential requirement to validate and enhance existing data.

The main stakeholder for the use case is the Snowdonia National Park Authority aiming, among other things, for the reduction of impact caused by invasive species. Therefore, three objectives are set:

- Increase the awareness of invasive and non-native species among the local population and tourists
- Engage citizens to monitor the occurrence, abundance and evolution of knotweed within the area to support distribution mapping
- Gather baseline information to inform management and implement a control strategy for knotweed

To facilitate the latter two, a strategy to assess the quality of crowdsourced observations is required to gain reliable information on the distribution of knotweed. This includes a validation procedure and fusion of observations with external data sources. The validation comprises thematic validation (Is it likely to be knotweed?) and spatial validation (Is the

Figure 3: Japanese Knotweed.



Photos taken by GBNNSS, Trevor Renals & Tube Lines Ltd. Obtained from http://www.nonnativespecies.org location valid and reliable?). Data fusion aims at the combination of an observation with other observations previously been made in that area and with authoritative data, in particular the modelled distribution map provided by Welsh Government. Accordingly, the fusion process supports both thematic and spatial validation procedures.

Input data for the use case, in particular the citizen observations, can be obtained from survey campaigns conducted by park authority staff, local residents and tourists using the COBWEB mobile application and Web-portal. Additional data can be harvested from third parties, such as environmental organizations and authorities, record centres or social media. As the cost of treatment for knotweed, including chemicals and human effort, is usually based on surface measures, polygon data is preferred over point data. The spatial accuracy stored with the observations depends on the acquisition method, primarily being GPS positioning and screen digitizing. If only a point is provided, an extent must be provided to indicate a certain expanse. Thematic properties should indicate the occurrence or absence of knotweed, including multimedia files to evidence an occurrence. The accuracy of this information mainly depends on the observer's knowledge base, which can be derived from professional status, training and/or observation activity patterns.

The prerequisites for the use case are at first the existence of a citizen engagement strategy to produce a certain number of available and accessible raw observations. Furthermore, spatial data fusion requires access to suitable reference data, in this case the authoritative knotweed distribution map. All services for QA and data fusion need to be set up and accessible, best via a standardized interface and well-defined inputs and outputs. An orchestration engine is required to set up the workflow as described before in an either pre-defined or ad hoc manner.

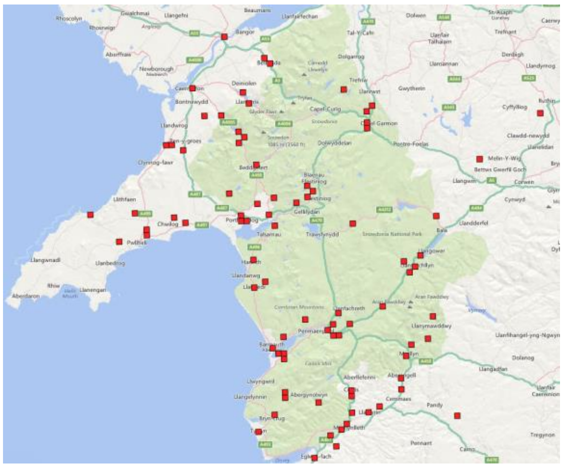
4 Prototype Implementation

To demonstrate the interaction between QA and data fusion workflows, a number of processes have been implemented and provided via OGC Web Processing Service (WPS) interface. Those are based on the 7 pillars of controls and checks described by [11] and provide metadata on:

- ISO19157 quality elements, such as positional, temporal and thematic accuracy, thematic correctness, usability and domain consistency,
- Elements for user capabilities, such as trust, judgement, validity and vagueness,
- Feedback quality elements similar to the approach taken by the GeoViQua project[1].

The combination of data fusion and data qualification services is realized through a BPMN (Business Process Modelling Notation) based workflow engine that composes HTTP (Hypertext Transfer Protocol) POST requests to WPS's and handles the responses as temporary datasets that can be reused in future processes, written as files, or pushed to a database via transactional OGC Web Feature Service (WFS). Post composition, workflows are saved as deployments that can be called via REST API or instantiated manually as a single run. The REST API lends itself to periodic qualification of records in a database via timed tasked such as Cron jobs. Further details on the BPMN workflow engine for WPS are described in [12].

Figure 4: Distribution of Japanese Knotweed in the Snowdonia National Park (1km grid cells with observed Knotweed presence, total area of SNP: 2.170 km²)



Source: NBN (https://data.nbn.org.uk) & Bing Maps

With regard to the knotweed use case described above, the current scenario comprises the following steps (Figure 5):

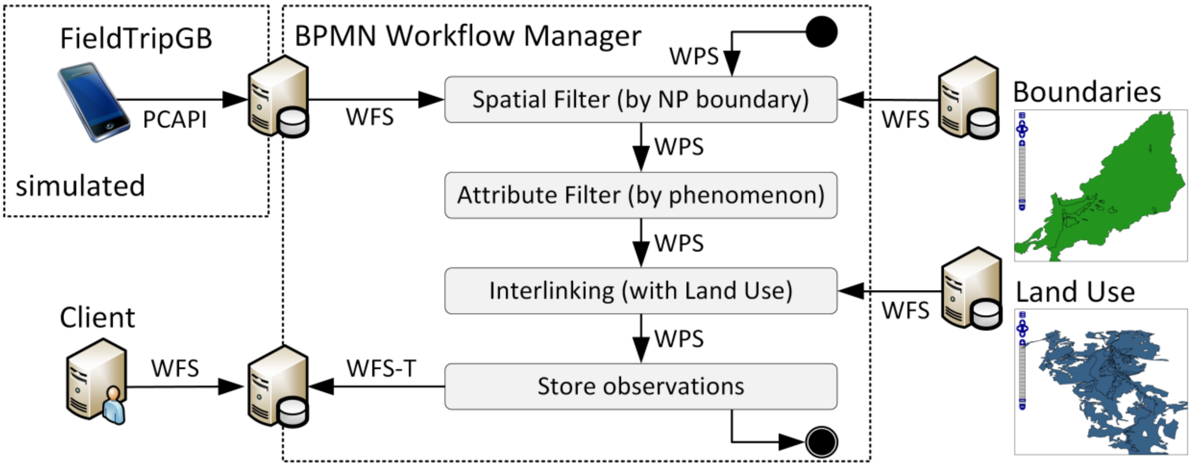
1. Observations are continuously made and uploaded as raw observations by the crowd. The data is stored in a repository with restricted access. Since there is no real data available yet, the observations are simulated for testing the workflow.
2. A spatial filter is applied to the observations using the national park boundaries obtained via WFS. The information, whether an observation is within the selected boundaries is stored with the observation.
3. Since observations on other species than knotweed are irrelevant to the use case, they are filtered accordingly using an attribute filter process.
4. To get contextual information on the observation, they are interlinked with land use data obtained via WFS. This information can be used to assess the potential impact of knotweed occurrence in that particular area.
5. Once the QA procedure is finished, the observations are written to the final repository via transactional WFS. The main attributes, beside geometry and location, as stored with the observations are shown in Table 1.

Once data passed the QA process, they can be accessed by clients for further analysis. In the knotweed use case, authorities can use the data as an indication of knotweed occurrences in the national park area. Even though field data is not yet available, the following mechanisms are introduced for further observation handling:

- Recent observations by different users increase the reliability of a knotweed sighting,
- Validation by experts, which can be either professionals or well-trained users, support the sighting and increase the reliability of the first observer,

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

Figure 5: Prototype workflow for quality assurance of crowdsourced observations for the knotweed use case



- Comparison with the existing knotweed distribution map can lead to the following results:
    - Observation of knotweed and corresponding occurrence on distribution map increase the probability rating of both the observation and the distribution map,
    - Observation of knotweed but no occurrence on distribution map is marked for expert validation,
    - Occurrence on distribution map but no observations available can be used to initiate targeted observations triggered at corresponding locations.

The comparison of observations with the knotweed distribution map is especially challenging, since the distribution map comes with its own uncertainty on the occurrence or absence of knotweed. The crowdsourced data may even be more accurate in some instances. Thus, regular expert validation for both the observations and distribution map is required to refine the classification process and produce reliable information on the distribution of knotweed in the Snowdonia National park, especially for rational decision and policy making.

5 Conclusion

In this paper we presented an approach for the combination of service-based QA and data fusion processes for the validation of crowdsourced observations. The selected use case is taken from the COBWEB project and deals with the distribution of Japanese Knotweed in the Snowdonia National Park, for which crowdsourced observations play an important role to identify and validate the occurrence of knotweed.

The strength of the approach lies in its flexibility and extensibility by using open data and interface standards for QA and fusion processes as well as for workflow management. Validation workflows can be built and customized in a toolbox approach using existing WPS processes that are organized and orchestrated using a BPMN workflow engine. Transferability is achieved by providing generic functionality for QA and data fusion that can be combined as required to serve an actual application. This adaption can be automated to a certain extent if information on inputs and target requirements for QA can be formalized.

Although as generic as standards will allow, the system requires endpoints with well-defined interfaces built on OGC standards to facilitate interoperability. To enable the operation of a generic workflow solution the service and target descriptions are required to be harmonized. The workflow engine does not actually contain any processing knowledge in itself, but relies on WPSs to act as process repositories that the workflow engine posts HTTP requests to process the data inputs.

To further increasing the range of functionality, additional processes will be added in further development. Moreover, a

Table 1: Main attributes for observations after running through prototype quality assurance

Observation attribute	Raw data value	Post-QA data value
fid	observation id	unchanged
userid	user id	unchanged
timestamp	observation timestamp	unchanged
species	species name	Japanese knotweed
capture	link to uploaded photos	unchanged
pos_acc	Positional accuracy of the observation (GPS accuracy)	unchanged
relation	-	Information on the spatial relation to land use data
DQ_Thematic	-	True, if observation of knotweed
DQ_SA_Relation	-	True, if within national park boundary

higher degree of automation is envisaged for service orchestration using well-defined and formalized process descriptions. On the non-technical side, privacy issues are always a concern with crowdsourced observations and demands for a proper authentication, authorization and access model. Furthermore, license restrictions need to be addressed in data fusion in a way that restricted data cannot be accessed by non-authorized users. Finally, a critical mass of observations need to be made in order to achieve robust and meaningful results.

References

[1] Clow, D., Makriyannis, E. (2011). iSpot Analysed: Participatory learning and reputation. In: Proceedings of the 1st International Conference on Learning Analytics and Knowledge. Banff, Canada. 34-43.

[2] de Jesus, J., Walker, P., Grant, M., Groom, S., 2012. WPS orchestration using the Taverna workbench: The eScience approach. Computers & Geosciences 47, 75–86.

[3] EA 2011. EAW Invasive Species Mapping, Japanese Knotweed 2011. Environment Agency, October 2011

[4] EA 2013. The knotweed code of practice: Managing Japanese knotweed on development sites (version 3, amended in 2013). Environment Agency, July 2013.

[5] Girres, J-F., Touya, G., 2010. Quality Assessment of the French OpenStreetMap Dataset. Transactions in GIS 14(4), 435-459

[6] Goodchild, M. F., & Li, L. (2012). Assuring the quality of volunteered geographic information. Spatial statistics, 1, 110-120.

[7] Haklay, M., 2010. How good is volunteered geographic information? A comparative study of OpenStreetMap and Ordnance Survey datasets. Environment and Planning B 37, 682-703.

[8] Hirth, M., Hoßfeld, T., Tran-Gia, P., 2012. Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms. Mathematical and Computer Modelling 57(11), 2918-2932.

[9] Hsueh, P. Y., Melville, P., & Sindhwani, V. (2009, June). Data quality from crowdsourcing: a study of annotation selection criteria. In Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing (pp. 27-35). Association for Computational Linguistics.

[10] ISO 2013. ISO 19157:2013 Geographic information - Data quality. International Organization for Standardization, ISO/TC 211.

[11] Meek, S., Jackson, M. and Leibovici, D., 2014. A flexible framework for assessing the quality of crowdsourced data. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, Spain.

[12] Meek, S., Jackson, M. and Leibovici, D., 2015. Addressing the quality assurance challenge for location-based crowd-sourced data through workflow composition of OGC web services. Computers & Geosciences (submitted).

[13] Safra, E., Kanza, Y., Sagiv, Y., Doytsher, Y., 2013. Ad hoc matching of vectorial road networks. International Journal of Geographical Information Science 27(1), 114-153.

[14] See, L. et al., 2013. Comparing the Quality of Crowdsourced Data Contributed by Expert and Non-Experts T. Preis, ed. PLoS ONE, 8(7), p.e69958.

[15] Song, W., Keller, J.M., Haithcoat, T.L., Davis, C.H., 2011. Relaxation-Based Point Feature Matching for Vector Map Conflation. Transactions in GIS 15(1), 43-60.

[16] Stadler, C., Lehmann, J., Höffner, K., Auer, S., 2012. LinkedGeoData: A Core for aWeb of Spatial Open Data. Semantic Web 3(4), 333-354.

[17] Veltkamp, R.C., Hagedoorn, M., 2001. State-of-the-art in shape matching. Principles of Visual Information Retrieval. Springer-Verlag, 87–119.

[18] Wang, F., Mäs, S., Reinhardt, W., & Kandawasvika, A. (2005). Ontology Based Quality Assurance for Mobile Data Acquisition.

[19] Wiemann, S. and Bernard, L., 2014. Linking crowdsourced observations with INSPIRE. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, Spain.

[20] Williams, F., Eschen, R., Harris, A., Djeddour, D., Pratt, C., Shaw, R.S., Varia, S., Lamontagne-Godwin, J., Thomas, S.E., Murphy, S.T. 2010. The Economic Cost of Invasive Non-Native Species on Great Britain. Technical Report, Centre for Agriculture and Biosciences International.

[21] Yang, B., Zhang, Y., Lu, F., Geometric-based approach for integrating VGI POIs and road networks. International Journal of Geographical Information Science 28(1), 126-147.

[22] Yuan, Y., Raubal, M., 2014. Measuring similarity of mobile phone user trajectories - a Spatio-temporal Edit Distance method. International Journal of Geographical Information Science 28(3), 496-520.