Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Linking crowdsourced observations with INSPIRE

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

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

Abstract

The combination of spatial data from the variety of sources on the web, being either legislative, commercially or voluntarily driven, is a major requirement for the establishment of a fully integrated geospatial web. Therefore, spatial data fusion techniques need to be linked to current web-developments, in particular on Spatial Data Infrastructures and the Semantic Web, to allow for standardized and effective use of combined spatial data for information retrieval. In this paper, crowdsourced environmental observations, representing the rapidly increasing amount of voluntarily collected data on the web, and INSPIRE, acting as the legal framework for spatial environmental information in Europe, are chosen to design and develop capabilities for spatial data fusion on the web. Possible use cases show mutual benefits for both volunteers and INSPIRE data providers, and might thus facilitate further collaboration. A prototypical implementation based on common SDI and Linked Data standards demonstrates the feasibility of the proposed approach and offers a starting point for further exploration.

Keywords: Data Fusion, SDI, Linked Data, Crowdsourcing, INSPIRE.

1 Introduction

The rapid development of location-enabled mobile devices and applications considerably influenced the public awareness, availability and use of spatial data in the recent years. Especially the amount of voluntarily collected spatial data on the web, often denoted as Volunteered Geographic Information (VGI, [4]), is continuously increasing. One of the most prominent examples is the OpenStreetMap[1] project for volunteered topographic mapping [5]. Furthermore, a large number of smaller projects, such as environmental monitoring campaigns [9], are establishing a versatile and most up-to-date basis for information retrieval and decision making that has not existed before.

In parallel, the INSPIRE (Infrastructure for Spatial Information in the European Community) regulation lays down the requirements for the unified and harmonized provision of spatial environmental data across the European Union und thus, serves as an ideal spatio-temporal reference for crowdsourced observations. In general, the integrated use of crowdsourced and administrative data will provide mutual benefits.

To link and combine arbitrary data sources on the web, service-based data fusion techniques need to be applied, independent from underlying data formats and provision means. Standards for information exchange and interlinking need to be established to support the interoperable and flexible orchestration of fusion processes and the effective combination of spatial data, one of the major building blocks towards a fully integrated geospatial web.

The paper introduces possible use cases for the fusion of crowdsourced environmental observations with INSPIRE reference data (chapter 2), followed by a short introduction to spatial data fusion (chapter 3). Subsequently, the requirements (chapter 4) and a prototypical implementation (chapter 5) for service-based spatial data fusion are described. Finally, a conclusion and outlook for further research is given (chapter 6).

2 Possible Use Cases

From the crowdsourcing perspective, INSPIRE can serve as a fundamental basis for validating environmental information on a broad range of topics, such as land cover and land use, environmental monitoring, habitat information or species distribution. However, INSPIRE will only be able to deliver coarse information with respect to spatial and (especially) temporal resolution, because of limited resources on the part of administrations responsible for data collection and provision. Thus, the use of INSPIRE data might be insufficient for data-intensive or real-time applications. The Eye on Earth[2] initiative already addresses this issue and tries to engage citizens to explore, collect and share environmental information in their surrounding and thereby complement existing data sources on the European level. However, developments to combine both crowdsourced and administrative data for information retrieval are still in the early stages. From the INSPIRE perspective, we identified

1 http://www.openstreetmap.org

2 http://www.eyeonearth.org

four different application use cases for the fusion with crowdsourced data:

1. *Data densification* can be applied to refine the response of a data service provider to a user request on a specific environmental phenomenon. At first, INSPIRE data is searched for information matching the request. If the data found is not yet sufficient in terms of spatio-temporal resolution or thematic attribution, crowdsourced data is selected and combined with the previous result to enhance the response accordingly. The data is provided to the user including lineage and quality information on the input and output data.
2. *Data enrichment* affects crowdsourced data that is not directly in the scope of the INSPIRE Annexes, but in any way connected to it. A user requesting environmental data automatically gets hints on related crowdsourced data, to indicate additional information sources. Moreover, relations to additional, explanatory or extending, data sources on the web facilitates the usability and applicability of INSPIRE data in general.
3. *Data update* describes the ability to refresh INSPIRE data with the help of related crowdsourced observations. Since data providers usually have limited resources for regularly updating information, crowdsourced information can give hints on where data actually needs to be updated. Therefore, both can systematically be compared to identify missing, incomplete, outdated or erroneous parts. The data provider is thus able to specifically investigate differences and, if applicable, update accordingly.
4. *Statistical analysis* of crowdsourced observations can benefit from the combination with INSPIRE, which acts as a persistent spatio-temporal reference system. Thereby, INSPIRE facilitates the comparability and ability to analyse crowdsourced information across the European Union.

All of the mentioned use cases require data fusion techniques to be applied, especially for matching, interlinking and resolving spatial data on the web. Furthermore, the awareness for lineage and quality information is considered crucial, because crowdsourced data will rarely comply with common INSPIRE quality standards.

3 State-of-the-art in data fusion

Within geosciences, the term fusion is quite ambiguous and frequently used within remote sensing, database research and spatial data processing [1]. With respect to signal processing, we hereinafter distinguish between data and sensor fusion. While sensor fusion is defined as the synthesis of multiple sensor measurements to receive comprehensive data on an observed phenomenon or entity [8], data fusion describes the combination of spatial data from multiple sources to provide a combined view that contains the most valuable data from the inputs [12]. Valuable hereby depends on the application context and purpose. Although sensor fusion is recognized as important for crowdsourced data, especially regarding consolidation and validation purposes, it is not in the scope of this paper. Instead, the focus is on data fusion, a task sometimes also referred to as conflation or data integration [10].

The classification of spatial data fusion is usually based on the operation direction (horizontal, vertical or temporal), input source (raster or vector), semantic level (representation, schema or ontology) or frequency (unique, periodic or real-time) [10, 11, 14]. The decision on a suitable fusion process depends on the application purpose, which can be change, discrepancy and error detection, data update and enrichment or the full integration of multiple spatial datasets [14]. Other factors describing a process can be supported in- and output formats, precision and recall rates or the computational performance.

To meet the requirements of web-based and standard compliant fusion of spatial data, a service-based approach is aspired. Therefore, complex fusion processes need to be decomposed into well-defined atomic processes in order to match the requirements of a Service Oriented Architecture (SOA) [3]. Here, we extend the previous approach presented in [13] and introduce the seven sub-processes shown in Figure 1 for the classification of service-based spatial data fusion. However, implementations might fall into more than one category and certain components may be optional, iterated or concatenated in different ways. Thus, the classification can be seen as an abstract reference framework for fusion processes.

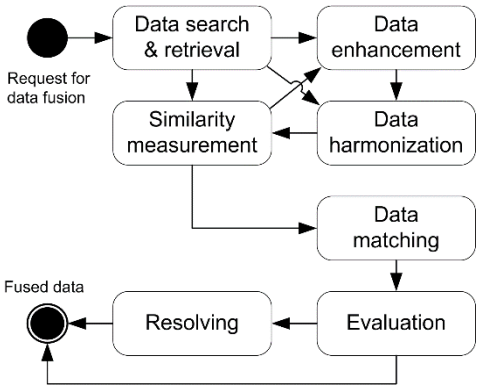
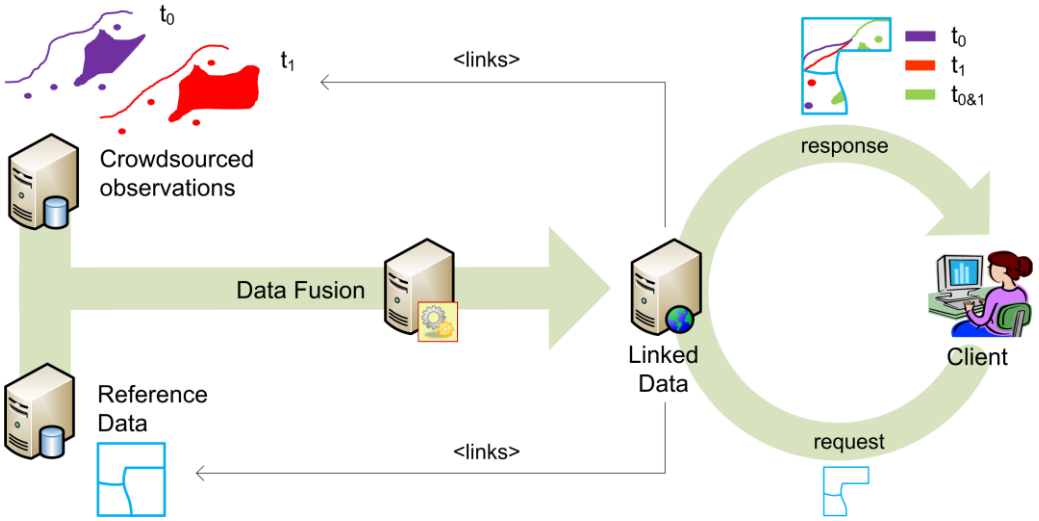


Figure 1: Classification for service-based spatial data fusion

4 Technical fusion aspects

For implementing web-based spatial data fusion, a number of service components for data provision, processing and search are required. For data provision and retrieval, INSPIRE specifies Data Download Services to be implemented for each data specification and member organization. The corresponding Technical Guidance currently obligates the use of the OGC (Open Geospatial Consortium) Web Feature

Figure 2: Generalized data fusion workflow for the combination of crowdsourced observations and spatial reference data



Service (WFS), but further specifications, such as the OGC Sensor Observation Service (SOS), the Web Coverage Service (WCS) or Linked Data are feasible [7]. For crowdsourced observations, the use of OGC services is desirable, but cannot be obliged. Communities often tend to develop their own data models and interface specifications to fit their specific purpose and thus, requires additional efforts on harmonization within the INSPIRE context. The applied services for spatial data processing should support the flexible and interoperable application and exchange of data fusion functionality across the web. Here, the use of the OGC Web Processing Service (WPS) is encouraged to comply with open standards. For the search of spatial data, the corresponding specification obligates the use of the OGC Catalogue Services for the Web (CSW) for INSPIRE [6]. For crowdsourced information, either the CSW or other open registries can be applied, as long as it can be accessed and requested online in a standardized manner.

The interlinking of crowdsourced observations with INSPIRE data can benefit from Semantic Web developments using Linked Data technology. Identified links can be encoded using RDF (Resource Description Framework) and either embedded directly in the crowdsourced data or managed as a standalone repository, which links to the corresponding data sources. Both approaches do not affect any INSPIRE infrastructure, but provide an additional layer for adding value to it. Beside the crowdsourced observations, further sources on the web, such as detailed descriptions on observed environmental phenomena by expert groups, can be linked as well. In addition, the service-based approach allows for the reuse of implemented data fusion functionality in other applications.

A generalized workflow for the fusion of crowdsourced and administrative data is depicted in Figure 2. Here, the starting point is a citizen collecting environmental information on a specific phenomenon in the field. This data is uploaded to a corresponding web portal, validated and stored accordingly. Depending on the application, this eventually triggers the fusion of the crowdsourced data with existing datasets, for which the selection is based on the observed phenomenon and its spatio-temporal extent. During the data fusion process, links between the input sources are generated and stored within a Linked Data triple store. Finally, a user requests and receives data on the phenomenon, including the option for added crowdsourced information. However, any of the use cases described in chapter 2 could be performed instead.

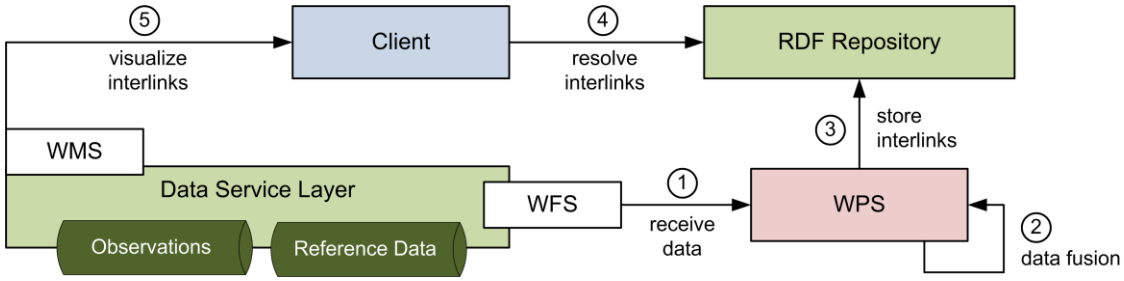
One of the most important aspects for data fusion as presented here, is the handling of quality information attached to the input sources, which determines the quality of a result and accordingly it's fit for purpose. Many applications or decision making processes require a certain level of data quality and thus rely on this information. To facilitate standardization and information sharing, quality associated to the input and output of a process should be formalized and encoded in a standardized format compliant to ISO 19115. While this is already mandatory for INSPIRE data, it is still a considerable challenge to add reliable quality information to crowdsourced data. The services for spatial data fusion must be capable of handling the quality information attached to the input sources. Hereby, the minimum requirement is to compile and keep track of the information from the input sources up to the final result. However, the optimum solution is a real utilization of quality information during the process, enabling propagated quality measurements based on input qualities and certain process characteristics.

5 Implementation prototype

The prototypical implementation for the fusion of crowdsourced environmental observations with administrative reference data is taken from the COBWEB project, which mainly deals with the development of on online-system for crowdsourcing in UNESCO Biosphere Reserves.

The current prototype manages the fusion of citizen observations on flora and fauna with INSPIRE relevant data on natural habitats. The basic workflow is depicted in Figure 3 and comprises the following steps:

Figure 3: Prototype workflow for the fusion of spatial data using OGC standards and a Linked Data repository



1. Crowdsourced observations and corresponding reference data is set up and provided via the OGC WFS interface for download. The fusion process, accessible via OGC WPS interface, requests this data as input for further processing.
2. The data fusion process is performed and relates the input features based on a number of similarity measurements, in particular bounding box overlap, Hausdorff distance and geometry buffer overlap for geometry objects and the Damerau-Levenshtein string distance for attribute comparisons.
3. All feature relations are encoded and stored as RDF triples including the corresponding WFS feature ids and underlying similarity measures.
4. The Client accesses the stored RDF triples and resolves the spatial features participating in a relation based on their feature id.
5. By using the OGC WMS interface on top of the data stores, the Client visualizes the resolved features by requesting a corresponding map overlay. A WMS GetFeatureInfo request can be generated to add detailed information on selected features.

The prototype currently follows a rather pragmatic approach to demonstrate the feasibility of the concept presented in this paper. To achieve the full potential of linking crowdsourced information to INSPIRE, further developments will focus on the support for all possible feature relations, the handling of quality information, automated orchestration of loosely-coupled fusion services with respect to the application use case and further interweaving the implementation with Semantic Web components.

6 Conclusion

So far, we demonstrated, that service-based spatial data fusion enables the combined use of multiple spatial data sources on the web for the retrieval of value-added information. Although still in the early stages, we can imagine a fully flexible, interoperable and versatile online system for dynamically interlinking and fusing any kind of spatial data on the web, with particular focus on the integration of the rapidly growing amount of voluntarily collected spatial data. Further developments and an evaluation of the approach will be documented in further publications.

The combination of INSPIRE data sources with data from crowdsourcing initiatives, offers great potential to create a comprehensive, most up-to-date and ubiquitously accessible source for environmental information in Europe. It combines the advantages of administrative data, namely quality assurance and the normative status, and crowdsourced data, with its rapid update cycle and partially high spatio-temporal resolution. The provided use cases show, that mutual benefits can be achieved from an advanced collaboration between both.

Still, one of the biggest challenges within the field of spatial data fusion remains the application-driven generation of value-added information from interlinked data. Therefore, it needs to be further analyzed how interlinked data can be selected and combined in an optimal fashion to serve a specific application purpose, and how quality information can be formalized and used to assist the fusion process. Although a lot of research questions still need to be solved, this will pave the way towards a Semantic Geospatial Web as proposed by Egenhofer [2].

References

[1] J. Bleiholder, F. Naumann. Data Fusion. *ACM Computing Surveys* 41(1):1-41, 2008.

[2] M. J. Egenhofer. Toward the semantic geospatial web. *Proceedings of the 10th ACM international symposium on Advances in geographic information systems*, McLean, USA, pp. 1-4, 2002.

[3] T. Erl. *SOA - Principles of Service Design*. Prentice Hall, 2008.

[4] M. F. Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal* 69(4):211-221, 2007.

[5] M. Haklay, P. Weber. OpenStreetMap: User-Generated Street Maps. *Pervasive Computing, IEEE* 7(4):12-18, 2008.

[6] INSPIRE. Technical Guidance for the implementation of INSPIRE Discovery Services. *Initial Operating Capability Task Force for Network Services*, Version 3.1, 2011.

[7] INSPIRE. Technical Guidance for the implementation of INSPIRE Download Services. *Initial Operating Capability Task Force for Network Services*, Version 3.1, 2013.

[8] OGC. OGC Fusion Standards Study Engineering Report. *OGC Public Engineering Report*. Open Geospatial Consortium, 2010

[9] H. E. Roy, M. J. O. Pocock, C. D. Preston, D. B. Roy, J. Savage, J. C. Tweddle, L. D. Robinson. Understanding Citizen Science & Environmental Monitoring. *Final Report on behalf of UK-EOF*. NERC Centre for Ecology & Hydrology and Natural History Museum, 2012.

[10] J. J. Ruiz, F. J. Ariza, M. A. Ureña, E. B. Blázquez. Digital map conflation: a review of the process and a proposal for classification. *International Journal of Geographical Information Science* 25(9):1439-1466, 2011

[11] A. Schwinn, J. Schelp. Design patterns for data integration. *Journal of Enterprise Information Management* 18(4):471-482, 2005.

[12] S. Stankuté, H. Asche. An Integrative Approach to Geospatial Data Fusion. *Computational Science and Its Applications – Proceedings of ICCSA 2009*, Suwon, Korea, pp. 490-504, 2009.

[13] S. Wiemann, L. Bernard. Conflation Services within Spatial Data Infrastructures. 13th AGILE International Conference on Geographic Information Science, Guimarães, Portugal. 2010.

[14] S. Yuan, C. Tao. Development of Conflation Components. *Proceedings of Geoinformatics'99 Conference*, Ann Arbor, USA, pp. 1-13, 1999.