Conflation Services within Spatial Data Infrastructures

Stefan Wiemann, Lars Bernard
Professorship for Geoinformation Systems, Technische Universität Dresden, Germany
{Stefan.Wiemann|Lars.Bernard}@tu-dresden.de

ABSTRACT

The conflation of spatial data is one important task concerning the generation of knowledge from available geoinformation. Research in this domain has been carried out since the early 80s and is still one of the most challenging parts within the field of spatial data processing. The range of tasks incorporates updating, change detection, enhancement and integration of spatial data. The current developments of *Spatial Data Infrastructures* (SDI) following the approach of *Service-oriented Architectures* (SOA) foster the development of distributed geospatial information services. The fusion of thematically comparable resources to enhance available spatial information is one of the core goals of SDI. This paper presents an approach towards ad-hoc conflations of distributed administrative and open source road data using OpenGIS compliant web services.

Keywords: Conflation, geoprocessing services, Spatial Data Infrastructure, data fusion

1. INTRODUCTION

In the field of geosciences the term Spatial Data Infrastructure has already been established. Lately, it is seen synonymous with the change of data-oriented towards service-oriented structures and comprises practical, legal, organizational, technical and social aspects (Wytzisk and Sliwinski 2004). SDI shall facilitate the cross-border cooperative use of distributed geoinformation services (Riecken et al. 2003), and thus promote the paradigm shift from previously dominant *Geographic Information Systems* (GIS) to open interoperable systems.

Characteristic problems of conventional spatial data handling are the significant number of different, often proprietary data formats and interfaces as well as heterogeneous metadata structures (McKee 2004). Furthermore the acquisition and utilization of spatial data is hindered by the existence of numerous data sets from various sources, since they differ in acquisition methods, data structures and attributions. Nevertheless, to create application-specific value-added information, spatial data fusion, in particular conflation is essential. It combines information from at least two spatial datasets to achieve enrichment in either the spatial or the attribute aspect (Yuan and Tao 1999). The matching of homologous features forms the core component of conflation and normally aims at unambiguous feature-mapping based on geometrical, topological and semantic attributes.

Today the possibilities of dynamically combining spatial information from different, distributed sources in SDI are still very limited. The need for conflation is caused by an increasing number of data provided through standardized interfaces and the objective of avoiding redundant information. In addition, on-going Web 2.0 developments and the appearance of small-sized mobile GPS receivers offer the possibility of user-generated spatial data beyond standardized SDI using citizens as "voluntary sensors" (Goodchild 2007). Therefore conformance and interoperability of spatial data and services play a major role in conflation processes.

To enable conflation within SDI, established methods must be combined with new service-oriented approaches. Hence, this paper proposes a service architecture for conflation and attempt to answer the following questions:

- What are the advantages of a service-oriented conflation in the context of SDI?

- To which extent can standardized metadata help to optimize and automate the process?
- Which sub-processes and interfaces are required to perform conflation within SDI?
- What are the requirements of an ad-hoc service-oriented conflation of spatial data?

The proof-of-concept implementation presented in this paper is restricted to the conflation of road data, since this is one of the most important applications concerning data fusion. It realizes a service-based approach on conflation based on existing standards. Two data sets from different sources are used to perform feature and attribute transfers and achieve value-added information. Furthermore a statistical analysis of the process support evaluation and improvement of services involved in the workflow and reduces the time required for interactive post-processing significantly.

2. STATE OF THE ART IN CONFLATION

Typically two kinds of conflation can be distinguished: horizontal and vertical. Horizontal conflation often refers to merging adjacent spatial data by *Zipping* methods (Beard and Chrisman 1986), whereas the present work deals with datasets covering the same area referred to as vertical conflation, more precisely the vertical conflation of vector data. The according fields of application are depicted in figure 1 and include updating, enrichment and integration of spatial data as well as the detection of changes (Yuan and Tao 1999).

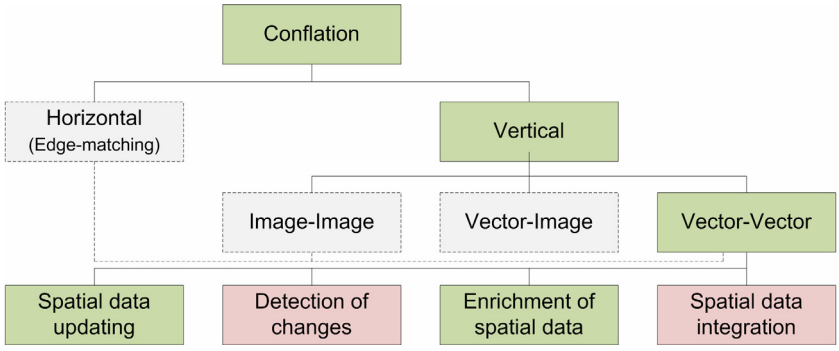

Figure 1: Conflation classification with main focus on vertical conflation of vector data.

Conflation consists of several sub-processes. At first required input data for conflation must be found, analyzed and compared to ensure suitability for further processing. Therefore geometric, thematic and structural properties, possibly derived from available metadata, are used. If the analysis of the considered datasets shows that considered data needs to be adjusted to allow the conflation processing, a pre-processing step is required. Pre-processing includes for instance map alignment and spatial or thematic generalization. This step is followed by first identifying and then assigning homologous objects to each other, the most important task towards successful conflation. Geometrical, topological and semantic attributes can be used to achieve an (ideally) unambiguous mapping. After the matching process value-added information can be created by joining assigned features and transferring required information.

Since the representation of same real-world entities in different spatial datasets are often very heterogeneous, feature matching is one of the biggest challenges in conflation. Conflicts like different coordinate reference systems, representation forms, resolutions or classifications must be solved to enable reliable matching. This is why most of the current approaches are realized as isolated applications with only a few links and not applicable in a generic manner to various data sets (Zhang et al. 2005). In general the matching process of vector data can be categorized in three ways (figure 2). First, based on the considered methods and attributes, matching can be divided into

geometric, topological and semantic matching. Secondly the processes can be categorized into bottom-up and top-down approaches, depending on either the sequence of node and edge matching or the sequence of semantic and geometric matching. The third differentiation of matching methods is based on the geometry type of the considered spatial data, in particular point, polyline and polygon. As various representations of the same real world entities typically differ from each other, similarity measurements must be established to assign homologous features. They can be distinguished by geometrical, topological and semantic criteria. However, the combination of the different measurements is the key factor for successful matching.

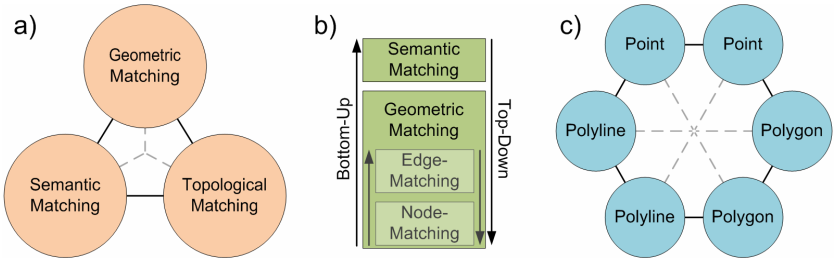

Figure 2: Differentiation of matching. As shown the processes can be distinguished by used methods and attributes (a), sequence of process (b) and geometry of input features (c).

Geometric similarity measurements are often based upon simple information like distance, length, angle or linearity (McMaster 1986). Furthermore combinations and indices, e.g. the Hausdorff-distance or Turning-Function, are used to describe and compare features to each other (Veltkamp and Hagedoorn 2001). Topological similarity is often combined with geometric measurement and depends on the relationship of features. Therefore approved approaches are the Spider-Function (Rosen and Saalfeld 1985) and the Round-Trip-Walk (Filin and Doytsher 2000), both depending on matched nodes and adjacent edges. Current research on similarity measurement often deals with semantic matching, usually based on thematic attributes, data structure or geo-ontologies (Weis and Naumann 2004, Giunchiglia et al. 2007).

3. CONCEPTUAL DESIGN OF SERVICE-ORIENTED CONFLATION

Conflation of provided spatial data can be realized by the introduction of the OGC *Web Processing Service* (WPS). The WPS interface offers geoprocessing capabilities within SDI and thus helps gradually providing interoperable and distributed processing and analysis services instead of monolithic GIS (Friis-Christensen et al. 2007). To achieve real benefit in comparison to established methods and implement complex processes, such as conflation, well defined service interfaces as well as performance and dynamic orchestration of services are crucial. The moving code paradigm, Grid Computing and stateful processing are most promising regarding the performance of web processing (Brauner et al. 2009). In addition transferred data can be reduced by using references, filters and caching methods, as network limitations, especially between clients and services, will influence the performance (Foerster and Schäffer 2007). Orchestration of services is essential for adding value to service oriented architectures and can be considered as creation of complex services by aggregating existing ones (Einspanier et al. 2003, Alameh 2003). XML-based BPEL (*Business Process Execution Language*) has emerged as the de-facto standard for web service orchestration (van der Aalst 2005) and has been already applied to the WPS interface (Schaeffer 2008). Furthermore cascading requests, the WPS standard itself and manual client-side aggregation can be used to build service chains for processing services (OGC 2007). Still the lack of semantics can be identified as the most challenging task concerning geoprocessing within SDI.

Service-based conflation of spatial data consists of several manually or dynamically chained data and processing services (figure 3). First of all suitable input datasets must be found. For this purpose OGC standards for catalogue services (CSW) and offering spatial data (e.g. WFS) should be used to ensure interoperability. The following comparison of input data can benefit from standardized metadata and results in a list of necessary pre-processing steps. This optional pre-processing for adjustment of spatial data can take advantage from existing services, such as transformation or generalization services. At this point, also a tiling of spatial datasets can be applied to facilitate asynchronous processing or GRID computing. The subsequent matching service depends on the input data and is based on a combination of geometrical, topological and semantic similarity measurements. The results of matching, e.g. a linkfile with specified schema documents, are thereafter used to join assigned features, from which value-added information can be generated by transferring information. Further analysis or evaluation might be done during an optional post-processing, which is followed by passing the result to the client via direct transfer or data service (e.g. WFS or WMS). The result will in general contain the conflated spatial data and in addition optional elements like extended information about input data, intermediate results or statistical analysis of sub-processes. Evaluation of the process is important to ensure a high quality of service and should be performed regularly.

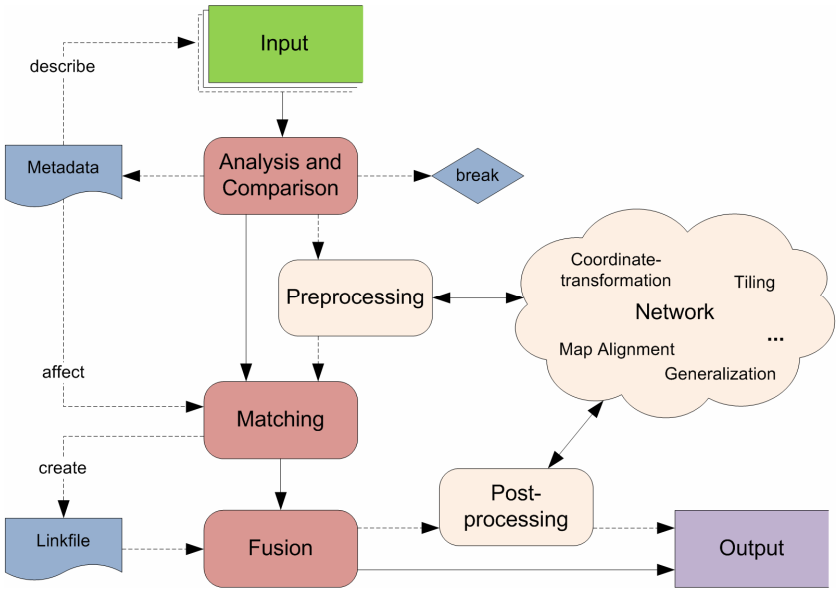

Figure 3: Workflow for service-based conflation within SDI

The search and binding of services involved in the workflow is enabled by catalogue services holding the corresponding metadata. Current research in this field is mostly related with semantic descriptions of spatial data and services. Here, the WPS standard proposes the utilization of WPS application profiles to uniquely indentify a process. An approach of how to use these profiles is presented by Nash (2008). Further research covers the combination of syntactic and semantic information (Lemmens et al. 2007) or the development of advanced ontology-based descriptions (Lutz 2007).

The combination of up-to-date spatial data within SDI is one of the most important advantages compared to conventional GIS-based methods. Thus, ad-hoc service-based generation of value-added information can support decision-makers in time-critical applications like risk or disaster management. Through current limitations in performance, this is especially applied on small datasets.

In addition server-side processing enables a thin client structure, so any portable device will be able to perform conflation of spatial data with only few requirements. The creation of flexible, efficient and reliable workflows strongly depends on quality and quantity of available services. Therefore both have to be considered when building a service-oriented infrastructure for conflation.

4. IMPLEMENTATION

Based on existing standards a service-based conflation has been implemented to evaluate how far ad-hoc data fusion can be realized in current SDI. Two road data services are used as an example to perform feature and attribute transfers and achieve value-added information. The first one is stemming from an administrative data source and follows the German ATKIS model (*German Authoritative Topographic Cartographic Information System*). The second data service uses data created on a voluntary basis from *OpenStreetMap* (OSM). Whereas the first data set is assumed to represent a typical reference data set with a well-known data schema, the second is considered as a less strictly structured but probably more up-to-date information source. Hence, in this example the ATKIS dataset is extended by adding features and attributes from OSM which did not exist before. Furthermore a statistical analysis is part of quality assurance for evaluation and improvement of services involved in the workflow. This analysis is mainly based on a weighted average of probabilities of single similarity measurements, which is saved as an additional attribute in the value-added dataset. In further processing this information supports linking homologous features and reduces the time required for interactive post-processing by customized visualization.

To build a proof-of-concept application for the designed SDI-based conflation approach, services for retrieval, processing and visualization of spatial data were implemented (figure 4). The software used are *Geoserver* (WFS and WMS) and the *52°North* WPS framework with Java-library *Geotools*. Service orchestration is realized by a simple Web-client, based on the server-side scripting language PHP and *libcurl* library.

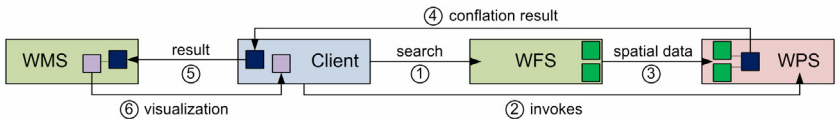

Figure 4: Simplified structure of the implemented conflation process

The processing services, core of the implementation, are divided into spatial data analysis and comparison, feature matching and the transfer of features and attributes. All intermediate results are saved in XML-structured files on the server and linked within the WPS response document. The visualization of conflation results is realized by using the REST API for *Geoserver* and JavaScript-based *OpenLayers* (figure 5). Applying Styled Layer Descriptor for WMS (WMS-SLD) feature matching probabilities can be visualized in different colors or grayscale, thus supporting error detection and interactive post-processing.

The average matching probability of features for the implemented services is about 80 to 90 percent. Most of the encountered problems during implementation can be traced back to heterogeneous structured input datasets. Especially OSM data often lacks topological consistency or spatial accuracy resulting from the collection of spatial data by volunteers partly having less experience and missing quality management procedures. This can cause incorrect matches and must therefore be corrected. Hence, a number of pre-processing steps or advanced similarity measurements need to be implemented to fulfill the expectations on high-value services for conflation. Still, results of the implementation are promising to future development and show that conflation approaches can be integrated in existing SDI with little effort.

Figure 5: Screenshot of the application. The matching probability is represented by grayscale (black indicates high probability). In addition transferred (long dashed) and not matched (short dashed) features are visualized.

5. CONCLUSION AND FURTHER RESEARCH

The developed concept and its implementation demonstrate a way to realize-oriented conflation in SDI and proof its feasibility. Relevant sub-processes and interfaces have been identified. Changing data sources or underlying software components does not affect the designed service interfaces, thus interoperability allowing for ease in use, efficiency, maintenance, and scalability could be achieved. In a next step the findings shall be fed back into standardization processes and ways to better performance of the ad-hoc conflation will be analyzed.

As semantics (metadata, standardized data specifications) play an important role in data fusion, the *Infrastructure for Spatial Information in the European Community* (INSPIRE) Directive is one of the driving forces in creating necessary basic conditions by promoting the publication of metadata standards and data specifications for European SDI. However, in the whole field of Geoinformatics there is a need for further research regarding the semantics of spatial data and services, also regarded as "one of the next frontiers in Geospatial Information Science" (Kiehle et al. 2006). In the future responses to queries for geospatial information shall not include irrelevant hits (Egenhofer 2002). This can only be achieved by open semantic reference systems and descriptions (Kuhn 2003) and fully interoperable systems. Also further development on performance and service chaining will contribute to automated, fully flexible, interoperable and user-friendly geoprocessing within SDI.

Current research is often related to governmental SDI. However, the development of Web 2.0 and the appearance of small-sized mobile GPS receivers offer the possibility of user-generated spatial data beyond standardized SDI. Different collection methods, data structures and attribution complicate the acquisition and utilization of spatial information from different sources. Therefore those community-based approaches should be considered when talking about interoperability of data and services. Although the integration of OSM within SDI will hardly be possible, both could benefit from each other.

The generation of knowledge from the variety of available spatial information through efficient processing, enrichment and use, play a decisive role in building knowledge-based structures. Since current approaches are often isolated applications and spatial data within SDI is typically very heterogeneous the development of generic algorithms for data fusion is crucial. In addition to the processing of vector data SDI offers various possibilities for conflation, such as real-time sensor data or coverages. Furthermore possibilities for service-based spatial data adjustment, like geometric or thematic generalization, have to be improved. Hence, service-based conflation and data fusion in general will be a major issue for future progress in the fields of spatial data processing and SDI-development.

BIBLIOGRAPHY

Alameh, N. (2003). Chaining Geographic Information Web Services. IEEE Internet Computing 7, September 2003, Nr. 5. 22–29.

Beard, M. K., Chrisman, N. R. (1986). Zipping: New Software for Merging Map Sheets. Proceeding of American Congress on Surveying and Mapping, 46th Annual Meeting. 153–161.

Brauner, J., Foerster, T., Schaeffer, B., Baranski, B. (2009). Towards a Research Agenda for Geoprocessing Services. 12th AGILE International Conference on Geographic Information Science, Hannover, June 2009.

Egenhofer, M. J. (2002). Toward the Semantic Geospatial Web. Proceedings of the 10th ACM international symposium on Advances in geographic information systems. 1–4.

Einspanier, U., Lutz, M., Senkler, K., Simonis, I., Sliwinski, A. (2003). Towards a Process Model for GI Service Composition. GI-Days, Münster, June 2003.

Filin, S., Doytsher, Y. (2000). The Detection of Corresponding Objects in a Linear-Based Map Conflation. Surveying and Land Information Systems, 60 (2000), Nr. 2. 117–128.

Foerster, T., Schäffer, B. (2007). A Client for Distributed Geo-processing on the Web. In Ware, J. M. (ed.), Taylor, G. E. (ed.). Web and Wireless Geographical Information Systems. Springer-Verlag, Berlin – Heidelberg. 252–263.

Friis-Christensen, A., Ostländer, N., Lutz, M., Bernard, L. (2007). Designing Service Architectures for Distributed Geoprocessing: Challenges and Future Directions. Transactions in GIS, 11 (2007). Nr. 6. 799-818.

Giunchiglia, F., Yatskevich, M., Shvaiko, P. (2007). Semantic Matching: Algorithms and Implementation. In Spaccapietra, S. (ed.). Journal on Data Semantics IX, Springer-Verlag, Berlin - Heidelberg. 1–38.

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. GeoJournal, 69 (2007), Nr. 4. 211–221.

Kiehle, C., Greve, K., Heier, C. (2007). Requirements for Next Generation Spatial Data Infrastructures: Standardized Web Based Geoprocessing and Web Service Orchestration. Transactions in GIS, 11 (2007), Nr. 6. 819–834.

Kuhn, W. (2003). Semantic Reference Systems. International Journal of Geographical Information Science, Guest Editorial, 17 (2003), Nr. 5. 405-409.

Lemmens, R., Wytzisk, A., de By, R., Granell, C., Gould, M., van Oosterom, P. (2007). Enhancing Geo-Service Chaining through Deep Service Descriptions. Transactions in GIS, 11 (2007), Nr. 6. 849–871.

Lutz, M. (2007). Ontology-Based Descriptions for Semantic Discovery and Composition of Geoprocessing Services. GeoInformatica, Volume 11, Nr. 1, March 2007. 1–36.

McKee, L. (2004). The Spatial Web. White paper, Open GIS Consortium, June 2004.

McMaster, R. B. (1986). A Statistical Analysis of Mathematical Measures for Linear Simplification. The American Cartographer, 13 (1986), Nr. 2. 103–116.

Nash, E. (2008). WPS Application Profiles for Generic and Specialised Processes. GI-Days, Münster. June 2008.

OGC (2007). Open Geospatial Consortium (ed.). OpenGIS Web Processing Service. June 2007, Version 1.0.0.

Riecken, J., Bernard, L., Portele, C., Remke, A. (2003). North-Rhine Westphalia: Building a Regional SDI in a Cross-Border Environment / Ad-Hoc Integration of SDIs: Lessons learnt. 9th EC-GI & GIS Workshop ESDI, June 25-27, 2003. Coruña, Spain.

Rosen, B., Saalfeld, A. (1985). Match criteria for automatic alignment. Proceedings of AUTOCARTO 7, March 1985. 456–462.

Schaeffer, B. (2008). Towards a Transactional Web Processing Service (WPS-T). GI-Days, Münster, June 2008.

Van der Aalst, W.M.P., Dumas, M., ter Hofstede, A.H.M., Russell, N., Verbeek, H.M.W., Wohed, P. (2005). Life After BPEL? In Bravetti, M. (ed.), Kloul, L. (ed.), Zavattaro, G. (ed.). Formal Techniques for Computer Systems and Business Processes. Springer-Verlag, Berlin – Heidelberg. 35–50.

Veltkamp, R. C.; Hagedoorn, M. (2001). State of the Art in Shape Matching. In Lew, M. S. (ed.). Principles of visual information retrieval. Springer Verlag, London. 87–119.

Weis, M., Naumann, F. (2004). Detecting Duplicate Objects in XML Documents. Proceedings of the 2004 international workshop on Information quality in information systems, Paris. 10–19.

Wytzisk, A., Sliwinski, A. (2004). Quo Vadis SDI? 7th AGILE Conference on Geographic Information Science, Heraklion, Greece. 43–49.

Yuan, Shuxin; Tao, C. (1999). Development of Conflation Components. Proceedings of Geoinformatics, Ann Arbor. 1–13.

Zhang, M., Shi, W., Meng, L. (2005). A Generic Matching Algorithm for Line Networks of Different Resolutions. 8th ICA Workshop on Generalisation and Multiple Representation, A Coruna, July 2005.