A framework for building multi-representation layers from OpenStreetMap data

M. Müller[1], S. Wiemann[1], B. Grafe[1]

[1]Technische Universität Dresden, Helmholtzstr. 10, 01062 Dresden
Email: {Matthias_Mueller | Stefan.Wiemann}@tu-dresden.de;{Bernd.Grafe}@mailbox.tu-dresden.de

1. Introduction

Within Geoscience the concept of Volunteered Geographic Information (VGI, Goodchild 2007) is currently on the rise. Among numerous projects in the field of VGI, OpenStreetMap (OSM) is one of the most advanced and has already been subject to several research studies[1]. Various secondary products and services are derived from OSM data[2].

Following the principle of a commons-based peer production (Benkler 2002) with participants acting as voluntary sensors (Goodchild 2007), OSM can serve as a source for geospatial modelling and analysis with global coverage. In contrast to classical Spatial Data Infrastructures (SDI) perspective, OSM data lacks standard conformity, provenance information, quality standards and the capability of complex feature modelling. However, the simplicity and flexibility of the OSM model seem to compensate these deficits. Furthermore, a number of research studies suggest a rather good quality in comparison with administrative or commercial datasets (Girres and Touya 2010, Haklay 2010, Ludwig et al. 2011).

All information collected by project participants is currently stored in one large database file containing raw feature data of the recorded spatial objects. The downside of one fine-grained feature data store is the rather big effort required for creating derived data such as thematic subsets, representations at different geographical scales or route networks. Useful complements – even for expert GIS (Geoinformation System) users – to the primary OSM database would be 1) a set of tools for data extraction, generalisation and processing and 2) a set of derived databases that contain refined data for particular fields of application. This motivates the service-based approach for deriving application–centric representations from OSM data presented in this paper.

The paper is organized as follows: The next chapter presents our general idea of a layered architecture for generating purpose-specific products from OSM data using service-oriented processing tools and multi-representation databases. Chapter three identifies tools which are required to create an application–centric OSM multi-representation database (MRDB). In chapter four we are presenting some preliminary findings on the creation of such a database for different types of OSM objects. The conclusion adds some ideas for building and maintaining a corresponding toolbox for OSM data.

[1] http://wiki.openstreetmap.org/wiki/Research (accessed: 09.08.2012)
[2] http://wiki.openstreetmap.org/wiki/List_of_OSM_based_Services (accessed: 09.08.2012)

2. An architecture for deriving application-centric representations from OSM data

The main purpose of the OSM project is to provide free to use and editable global maps under an open data license. Since its initiation in 2004 the mostly voluntary users contributed an enormous amount of data to its database. The application fields for using OSM data are manifold. Just as manifold are the requirements for data preparation, such as for:

- Map creation which requires model and cartographic generalisation methods for displaying objects at different scales as well as symbolisation rules for map rendering,
- Routing which requires the generation of a topological network including specific attributes like turns, impedances or speed limits,
- Location-based Services which require information on features of interest including thematic attributes and semantic descriptions.

Although each task may have its own data preparation needs, a number of generic tools can serve a multitude of applications. Those should be interchangeable and interoperable in order to build dynamic, reproducible and flexible workflows for deriving different representations of the primary OSM database. We suggest a service-oriented approach that facilitates the generation of derived and possibly simplified representations for OSM data. Such layers may then be used to support a variety of end-user services and applications.

As depicted in Figure 1a), the planet OSM database resides at the lowest level providing a uniform resource of data for the upper level components. The processing resides directly on top of this layer and provides data manipulation functionality like generalisation, transformation or symbolisation for data that is compatible to the original OSM data schema. The two upper levels contain application specific data and services. The multi-representation layer contains secondary data strongly related to end-user services in the product layer. This layered approach allows for better decoupling of OSM-related products from the rapidly growing database but does not foresee that users create their desired products on demand. However, the concept is capable enough to capture more sophisticated use cases where experts access to the processing layer directly to produce customized representations from the OSM database (Figure 1b). These representations are probably purpose-specific and for in-house or private use so there is little need to immediately publish them in an MRDB layer. However, if there should emerge a public demand for a particular representation it can be easily instantiated in the MRDB layer.

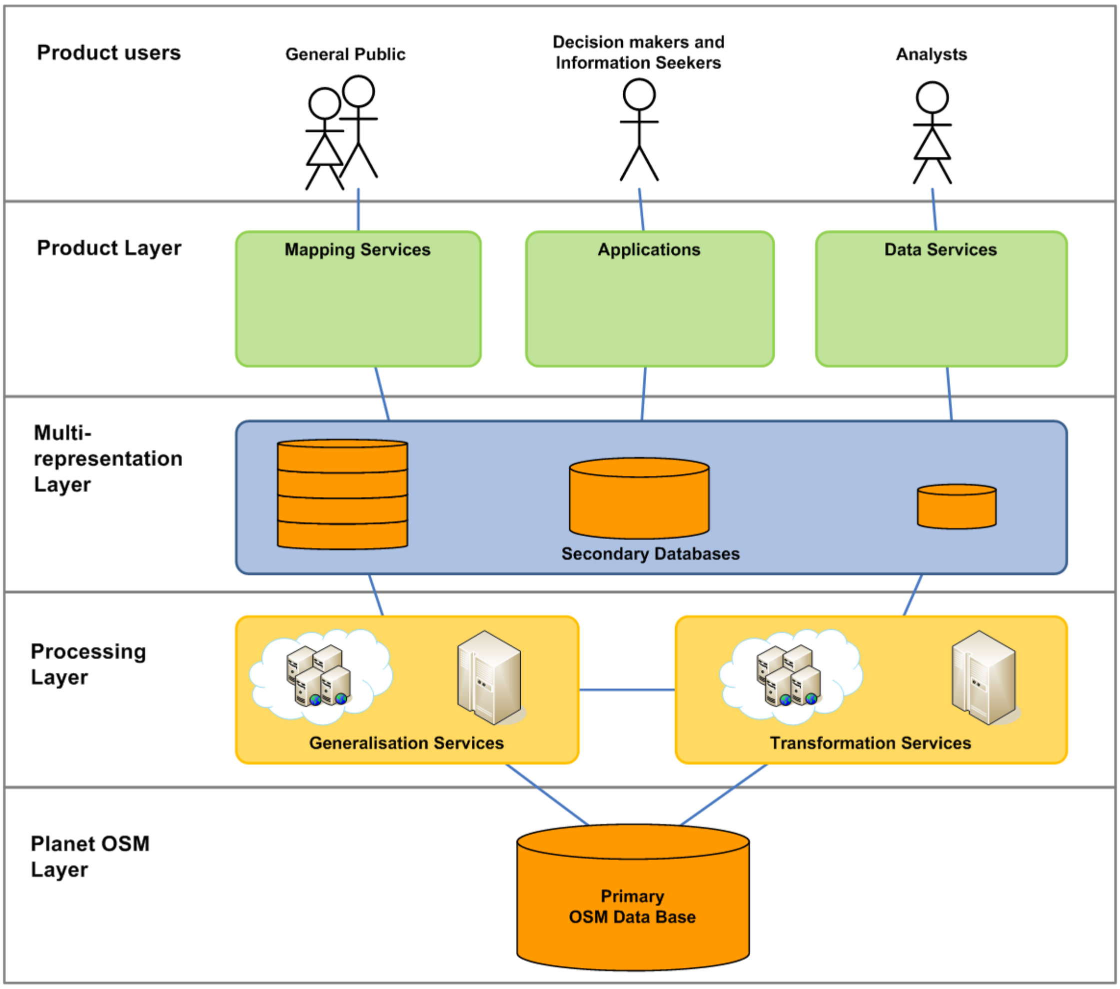

Figure 1a). A layered architecture for generating application-specific representations of the primary OSM database.

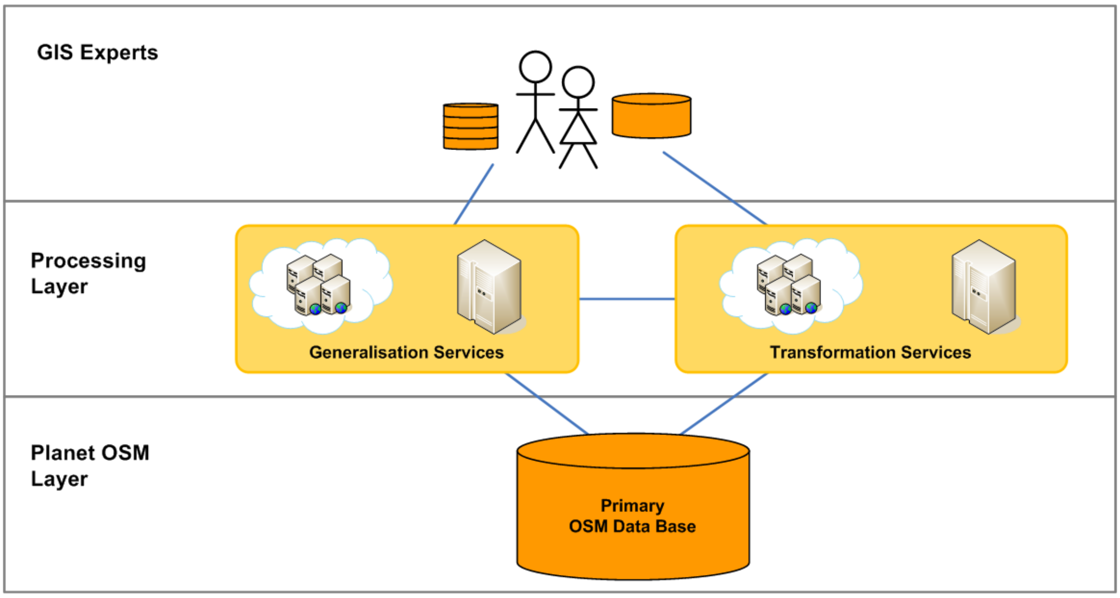

Figure 1b). Proposed architecture for generating customized representations of the primary OSM database.

3. Building a multi-representation layer for OSM data

In the literature, several approaches for building Multi Representation Databases can be found, focussing on the development of specific database schemata (Jones et al. 1996, van Oosterom 2005), the modelling of relationships within the database (Bobzien et al. 2008) or methods for deriving multiple feature representations (Devogele et al. 1996, Hampe et al. 2003, Brewer et al. 2007). However, within a service-based environment, the following requirements must be met:

- Workflows are decomposable into a set of atomic tools to facilitate flexibility and software reuse
- Functionality is decoupled from the underlying application and provided via open standardised interfaces to facilitate interoperability
- Tools provide sufficient syntactic and semantic descriptions via their interfaces to enable cataloguing and workflow enactment

Existing approaches for the creation of MRDBs need to be analysed and reworked with regard to the above mentioned requirements. Although this is expected to be a lot of work, it would significantly improve accessibility and utilisation of previous work in this field and perfectly match the open access policy for scientific research promoted by the European Commission.

To build an MRDB for OSM data in a flexible and reusable manner, the following steps need to be covered, ideally by a multitude of tools:

- Database creation – comprises the database schema creation including the different representation layers for each scale as well as the initial filling of the database using the global OSM database. Different parameters could be used to define spatial subsets, object filters or representation layer scales. Tools might benefit from existing semantic links between OSM features such as provided by the LinkedGeoData[3] project, encoded using the Resource Description Framework (RDF). Furthermore, rules for the creation of the MRDB could be formalized as presented by Brewer et al. (2007).
- Data harmonisation – required to compensate the different sources and acquisition methods of OSM data, which inevitably lead to different levels of detail in geometry and attribution throughout the whole database.
- Database Generalisation – includes primarily model generalisation processes to create suitable object representations based on previously defined layer scales. Different OSM objects require different generalisation strategies based on feature type, application area and scale. An overview on the structure and categorisation of web-based generalisation services can be found in Burghardt et al. (2005). A formal classification of available generalisation operators is given by Foerster et al. (2007). This point could be seen as an extension to the work on Web Generalisation Services previously presented by Foerster et al. (2008).
- Feature Matching – provides utilities to establish links between different scales in the MRDB. A classification of feature matching approaches is provided by Ruiz et al. (2011). A prototypical service based implementation using OSM data is described by Wiemann and Bernard (2010). The support for robust heuristics such as the strokes concept (Thomson and Brooks 2002, Thomson 2003) might further advance both automated generalisation and feature matching processes.
- Map rendering – includes cartographic symbolisation and generalisation to produce maps at different scales from the MRDB using standardised interfaces such as the Web Map Service (WMS) developed by the Open Geospatial

[3] http://linkedgeodata.org (accessed: 09.08.2012)

Consortium. On-the-fly rendering could be implemented by applying either tile-caching as presented by Goetz et al. (2012) or specialized database structures like the TGAP structure presented by van Oosterom (2005). In addition, sophisticated map generalisation processes as described by Neun et al. (2009) or the application of cartographic web services (Iosifescu et al. 2009) are promising approaches towards a toolbox for web based map rendering.

4. Implementation of an MRDB for OSM data

A challenging use case for the suggested concept is the provision of a capable Web Map Service for OSM data. To deal with the large amount of data and the data-intensive rendering process many existing mapping applications rely on tile map services and web caches. These approaches have well-known drawbacks in terms of storage resources to hold large volumes of tiles on the lower scales and are less suitable for rendering many customized styles, e.g. by the application of Styled Layer Descriptors (SLD).

Our aim is to create a pre-generalized database that is suitable for a range of map-rendering applications and to provide and sustain a transformation and generalisation toolbox for the OSM universe. Figure 3 shows a general workflow for such a data reduction task. By sub-setting and filtering the original OSM database can a reduced to the number of object and attribute data that are effectively used by the rendering stage. A subsequently applied generalisation procedure removes the excess geometric detail which will be anyway discarded by the renderer.

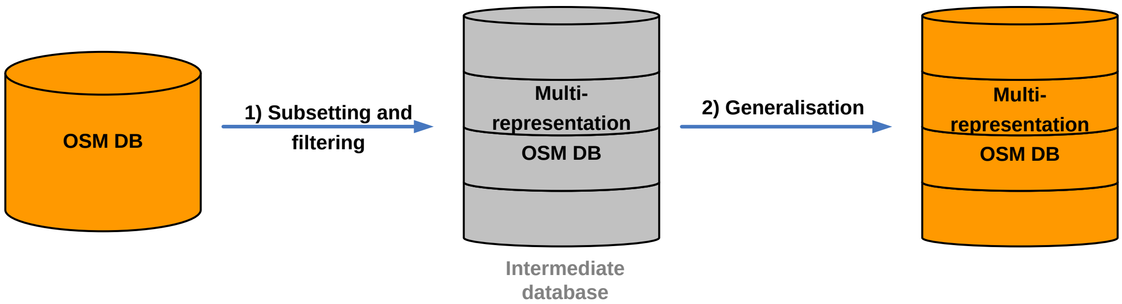

Figure 3. A simple 2-step workflow for creating an MRDB for OSM data.

At first, an initial MRDB is built by sub-setting and filtering the primary OSM database for each level of representation. These levels correspond directly to the scale ranges for map portrayal. We have tried this procedure using Mapnik styling and filtering rules to create suitable SQL sub-setting statements. The choice of these rules is arbitrary and any machine-readable rule format can be used. Due to the maturity and practical relevance of Mapnik we are considering its rule set as a good starting point.

In a second step each of the intermediate representations is generalized to the appropriate level of detail. The OSM database contains three types of geometries which are all relevant to the rendering process: polygons, lines and points. Altogether there is a plethora of algorithms to generalise these geometries and there is no “right” procedure to accomplish this task. Currently we are exploring a range of techniques that are able to yield a good amount of data reduction while being able to quickly and robustly process fairly large volumes of data (Figure 4). At least for medium to small scales, raster-based algorithms are scaling very well with the number of features. This suggests that an increasing amount of buildings’ footprints in the primary database will not affect the efficiency of these methods. With larger scales it may be necessary to apply more sophisticated algorithms to obtain building blocks for dense settlements rather than individual buildings. For such algorithms (e.g. Glander and Döllner 2007) a heavier computational burden can be expected with a significantly growing number of buildings' footprints.

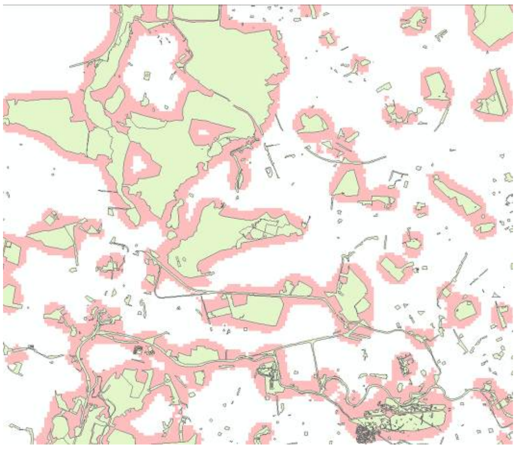


Figure 4. Preliminary generalisation results for polygon data. Green areas represent the original geometries; red areas represent the expanded (first image) or condensed areas (second image). Pictures are adjusted to a presentation scale of 1:300.000.

While polygon data can be handled satisfyingly by density-based procedures or morphological operations in the raster domain, line features – especially roads – are still an issue. We found that OSM data contains a vast number of tiny road segments instead of complete streets. Simply applying a geometric generalisation algorithm to each of these segments does not yield much data reduction.
Consequently, prior to geometric generalisation, a road reconstruction step is required where individual segments are recomposed to complete road geometries. However, the naive solution of merging adjacent segments with the same thematic attributes will not work in many cases because related road segments may have been collected by different people with different quality. Thus the attributes of neighbouring road segments may strongly vary, making it hard to determine all segments that form a road in the real world. Here we require robust heuristics that make reasonable assumptions about the similarity of adjacent segments to reconstruct the road features and to finally build a road network. One such candidate is the strokes concept (Thomson and Brooks 2002, Thomson 2003). A stroke is defined as a set of one more segments in a non-branching connected chain. In contrast to a merged geometry, the construction process for strokes also incorporates geometric properties such as line continuity, e.g. determined by the (minimal) deflection angle at junctions. This additional may help to identify a connecting road segment in the case of missing or ambiguous attribute information, i.e. a missing type. Once completed the algorithm returns a hierarchical road network, thus making an ideal starting point for further generalisation.
The actual generalisation task requires algorithms that give good control about the level of accuracy while providing good vertex reduction at the same time (e.g. Imai and Iri 1986). With a growing road network in the OSM database the generalisation procedure becomes more time consuming with every additional feature. To save computation time the different levels of generalisation should preferably build upon each other. For the architecture from chapter 2 this means that the multi-representation layer and the tool layer become more intertwined: Data reduction tools should not only

work for the primary database but also with the derived databases. Here compatibility can be achieved best if the representations of the original data generally stick to a strict subset of the original OSM data model.

Point objects currently do not require pre-generalisation for quick display. They are only displayed few at a time at very large scales and are thus completely removed form medium and small scale representations. However if the number of points will grow by orders of magnitude it may be convenient to generalise them in hot spot areas by either additional filtering or density reduction.

5. Conclusion and further developments

This paper has presented an architectural approach to generate and provide application–centric representations of the OSM database. The architecture introduces additional services on top of the primary database to facilitate the creation of OSM-based applications. Outsourcing data management and generalisation tasks into separate architectural layers reduces the implementation burden for OSM-based applications. For reusability we have introduced a multi-representation layer that contains reduced versions of the primary OSM database. To preserve flexibility the data transformation and data reduction tools are conceptualised as services in a processing layer and can thus be reused in various contexts and invoked for the creation of new representations.

A mapping use case has illustrated issues with existing approaches on OSM data handling and sketched the potential benefits of derived OSM databases, e.g. for mapping tasks, and the advantage of reusable generalisation tools for OSM data. An investigation of generalisation problems for different geometric objects in the OSM database discusses current and upcoming issues related to the characteristics and the volume of the data. Here the provision of both effective and efficient services for road generalisation has been identified as a pressing need to create simplified (and visually appealing) representations for large to medium scales.

An open issue is the communication of the quality of the derived representation compared to the primary OSM database. The generation of a simplified representation is not a real-time task and, depending on the applied algorithms and transformations, a time and resource consuming procedure. Constraints in terms of processing time and geometric quality of the simplified data require quality information to be associated with all derived representations. In many cases the derived representation will not adapt the daily update schedule of the primary OSM database. To avoid the use of outdated representations the quality information must also comprise the update schedule and link back to the used version of the primary OSM database.

Committing resources for computational intensive tasks for the provision of Web services for data generalisation can hardly be achieved by the OSM community. Nevertheless it is desirable to have the resources in the processing layer offered in a service-oriented manner.

The Moving Code concept (Müller et al. 2012, 2010) provides means to exchange computational logic in a similar way to exchanging data. Moving Code packages are self-describing service-oriented software components and may be used to deliver a variety of processing tools to different processing locations. The description of the contained code comprises 1) the interface description using Web Service standards, 2) a structured description of the required runtime environment, 3) a structured description of hardware requirements according to the OCCI (Open Cloud Computing Interface) standard, and 4) licensing information preferably by using the well-established Creative Commons licensing scheme. This information is packaged with the actual implementation (the “code”). Once downloaded the contained code can be immediately executed locally or deployed on a Web Server. Exchanging processing tools instead of offering public services for resource-hungry processing tasks is also more cost-efficient from a provider perspective. The envisaged toolbox can be hosted by any institution without overextending their resources. This makes the concept an ideal means to create an extensible generalisation and transformation toolbox for OSM products which can be shared with the community. Furthermore, it allows wrapping and reusing existing generalisation tools such as those provided by the WebGen framework[4].

Acknowledgements

We would like to thank the anonymous reviewer who helped to improve the quality of this paper.

References

Benkler Y, 2002, Coase’s Penguin, or, Linux and The Nature of the Firm. *The Yale Law Journal,* 112(3): 369–446.

Bobzien M, Burghardt D, Petzold I, Neun M, Weibel R, 2008, Multi-representation Databases with Explicitly Modeled Horizontal, Vertical, and Update Relations. *Cartography and Geographic Information Science,* 35(1): 3–16.

Brewer CA, Buttenfield BP, Frye C, Acosta J, 2007, SCALEMASTER: Multi-Scale Mapmaking from Multiple Database Resolutions and for Multiple Map Purposes. XXIII International Cartographic Conference (ICC2007), Moscow, Russia.

Burghardt D, Neun M, Weibel R, 2005, Generalization Services on the Web Classification and an Initial Prototype Implementation. *Cartography and Geographic Information Science,* 32(4): 257–268.

Devogele T, Trevisan J, Raynal L, 1996, Building a Multi-Scale Database with Scale-Transition Relationships. Advances in GIS II (Proceedings SDH ‘96), London, U.K.

Foerster T, Stoter J, Köbben B, 2007, Towards a formal classification of generalization operators, Proceedings of the 23rd International Cartographic Conference (ICC), Moscow.

Foerster T, Burghard D, Neun M, Regnauld N, Swan J, Weibel R, 2008, Towards an Interoperable Web Generalisation Services Framework – Current Work in Progress. Proceedings 11th ICA Workshop on Generalisation and Multiple Representation, Montpellier, France.

Girres J-F and Touya G, 2010, Quality Assessment of the French OpenStreetMap Dataset. *Transactions in GIS,* 14(4): 435–459.

Glander, T and Döllner, J, 2007, Cell-Based Generalization of 3D Building Groups with Outlier Management. 15th International Symposium on Advances in Geographic Information Systems (ACM GIS).

Goetz M, Lauer J and Auer M, 2012, An Algorithm Based Methodology for the Creation of a Regularly Updated Global Online Map Derived From Volunteered Geographic Information, GEOProcessing: The Fourth International Conference on Advanced Geographic Information Systems, Applications and Services, Valencia, Spain.

Goodchild MF, 2007, Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0. *International Journal of Spatial Data Infrastructures Research,* 2(2007): 24–32.

Haklay M, 2010, How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design,* 37(4): 682–703.

Hampe M, Anders K-H, Sester M, 2003, MRDB Applications for Database Revision and Real-Time Generalisation. Proceedings of 21st International Cartographic Conference, Durban/South, Africa.

Imai, H and Iri, M, 1986, Computational-geometric methods for polygonal approximations of a curve. *Computer Vision, Graphics, and Image Processing,* 36 (1986): 31–41.

Iosifescu I, Hugentobler M, Hurni L, 2009, Cartographic Web Services and Cartographic Rules – a new Approach for Web Cartography. Proceedings of the 24th International Cartographic Conference (ICC), Santiago, Chile.

[4] http://kartographie.geo.tu-dresden.de/webgen_wps/

Jones CB, Kidner DB, Luo LQ, Bundy GLl, Ware JM, 1996, Database design for a multi-scale spatial information system. *International Journal of Geographical Information Systems.* 10(8): 901–920

Ludwig I, Voss A and Krause-Traudes M, 2011, A Comparison of the Street Networks of Navteq and OSM in Germany, *Advancing Geoinformation Science for a Changing World,* Springer Berlin / Heidelberg, 65–84.

Müller, M, Bernard, L, Brauner, J, 2010, Moving Code in Spatial Data Infrastructures – Web Service Based Deployment of Geoprocessing Algorithms. *Transactions in GIS*, 14(S1): 101–118.

Müller, M, Kadner, D, Bernard, L, 2012, Moving Code – Sharing Geospatial Computation Logic on the Web. 15th AGILE International Conference on Geographic Information Sciences (AGILE 2012), Avignon.

Neun M, Burghardt D, Weibel R, 2009, Automated processing for map generalization using web services. *Geoinformatica*, 13, 425–452.

Ruiz JJ, Ariza FJ, Ureña MA, Blázquez EB, 2011, Digital map conflation: a review of the process and a proposal for classification. *International Journal of Geographical Information Science*, 25(9): 1439–1466.

Thomson, RC, 2003, Bending the axial line: Smoothly continuous road centre-line segments as a basis for road network analysis. Proceedings of the Fourth International Space Syntax Symposium, London.

Thomson, RC and Brooks, R, 2002 Exploiting Perceptual Grouping for Map Analysis, Understanding and Generalization: The Case of Road and River Networks. *Graphics Recognition Algorithms and Applications*, Springer Berlin / Heidelberg, 148–157.

van Oosterom P, 2005, Variable-scale Topological Data Structures Suitable for Progressive Data Transfer: The GAP-face Tree and GAP-edge Forest, *Cartography and Geographic Information Science,* 32(4): 331–346.

Wiemann S, Bernard, L, 2010, Conflation Services within Spatial Data Infrastructures. 13th AGILE International Conference on Geographic Information Sciences (AGILE 2010), Guimarães.