LINKED DATA - A MULTIPLE REPRESENTATION DATABASE AT WEB SCALE?

Stefan Hahmann and Dirk Burghardt

Stefan.Hahmann@tu-dresden.de, Dirk.Burghardt@tu-dresden.de

Dresden University of Technology, Institute for Cartography, Helmholtzstraße 10, 01069 Dresden

ABSTRACT

Recent developments in cartographic applications employ methods of the Web 2.0 and an increasing amount of user generated geoinformation. This leads to a considerable amount of up to date but heterogeneous data. Development of methods for interoperability on semantic level is required to use these data sources. Therefore an introduction to the (Spatial) Semantic Web and prototype applications will be presented. With the growing amount of information being moved from classical databases to the web there is a paradigm shift from the web of documents to the web of data. The LinkedGeoData project as an RDF implementation of the OpenStreetMap dataset has the capability to serve as a central interlinking hub for geodata within the Semantic Web. Linking of different information and representations of the same object is a precondition for queries that include different features. To discuss the question, whether Linked Data and the Semantic Web can serve as a Multiple Representation Database (MRDB) at web scale similarities and differences between both concepts will be shown. As an outlook ongoing work on connecting LinkedGeoData and Geonames with the aim of validation and enrichment of community data will be shortly described.

Keywords: *Semantic Web, Spatial Semantic Web, Linked Data, MRDB, Volunteered Geographic Information, Web 2.0*

INTRODUCTION

During the first decade of the 21st century the Internet community rapidly evolved the technologies of the "Web 2.0" (O'Reilly 2005). These technologies allowed users to

easily contribute data to the Internet. Information flow developed from a *unidirectional producer to user* pattern to a *bidirectional producer to user and user to producer* pattern. Thus user und producer, regardless if expert or amateur, were transformed to "*produser*" of geodata (Budhathoki et al. 2008). The first major application of the Web 2.0 was Wikipedia[1] followed by platforms such as youtube[2], blogs and social networks. Meanwhile huge amounts of collaboratively collected data can be found on the web. For the geographic domain OpenStreetMap[3] (OSM), which started in 2004, has a massive impact, because it is the first global, comprehensive and accessible source of geoinformation, which can be used free of charge and free of license restrictions. But also GeoNames[4] , Flickr[5] or even Wikipedia can now be used as geographic information sources. All these projects collect "volunteered geographic information (VGI)" (Goodchild 2007). The DBpedia mobile project (Becker 2008) is a use case, which demonstrates the application of VGI combined with semantic web technologies. Its location-aware client uses data of Wikipedia and Flickr to enrich an OSM background map with points of interests in multiple languages.

Due to low cost GPS receivers, a network of "human sensors" (Goodchild 2007) has been equipped with a tool to collect VGI and due to Google Earth, Google Maps and public Web Map Services this network has been enabled to enrich their collected spatial data with information that could be gathered from different kinds of satellite and aerial images.

SERENDIPITY - REUSING DATA BY LINKING DATA

Together all Web 2.0 applications provide a comprehensive information source. Linking this information and making them meaningful to computers to allow automatic processing and reasoning has a big potential to generate new knowledge from interconnected information sources. "It is the unexpected re-use of information which is the value added by the web" (Berners-Lee 2006). The fact of accidentally finding information, which is important for some purpose, while looking for something entirely unrelated is also known as serendipity (Wikipedia 2010).

[1] http://wikipedia.org/

[2] http://www.youtube.com/

[3] http://www.openstreetmap.org/

[4] http://www.geonames.org/

[5] http://www.flickr.com/

There already exist approaches for the geometric linking of data by finding corresponding objects in different point based (Beeri et al. 2005, Samal et al. 2004), polyline based (Samal et al. 2004) and polygon based (von Gösseln and Sester 2004) data sets. However, the bigger challenge seems to achieve semantic interoperability across multiple ontologies by aligning different ontologies (van Harmelen 2008).

THE SEMANTIC WEB

In 1998 Tim Berners-Lee, one of the inventors of the Internet and today a director of the W3C, introduced the principles of the "Semantic Web" (Berners-Lee 1998a, Berners-Lee 1999, Berners-Lee 2006). The *Semantic Web* aims to make the *World Wide Web* that was initially made for human consumption intelligible not only to humans but also to machines. Though the web is "machine-readable" it is not "machine-understandable" (Lassila and Swick 1999).

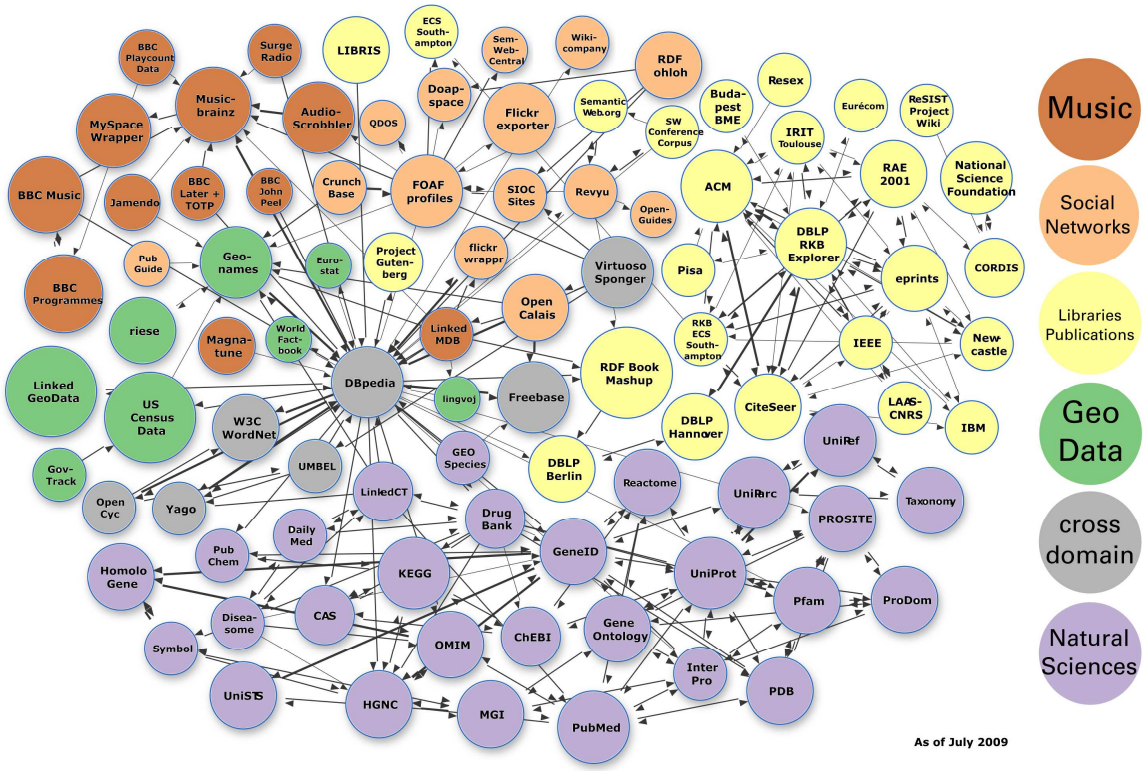

Fig. 1 The Linked Open Data dataset cloud of the Semantic Web. Modified after (Cyganiak and Jentzsch 2010)

For the Semantic Web the Resource Description Framework (RDF) is the core technology. RDF allows data providers to publish data and specify the semantics of the data in an interoperable way in the World Wide Web (Lassila and Swick 1999). Once

RDF data is published on the web and linked to other data sources, this data is called linked data. The Linked Open Data[6] project aims to be a central point for at least all openly accessible linked data which is published on the Semantic Web (*Fig. 1*). It should be emphasized, that a significant part of the linked open data cloud is geo related.

The semantic description of data can be accomplished by using vocabularies, as presented in the RDF Schema recommendation by the W3C (Brickley and Guha 2004). An extension of the RDF Schema recommendation is the Web Ontology Language (OWL) as presented in (van Harmelen 2008).

The data model of the Semantic Web is similar to the data model of relational databases (RDB) (Berners-Lee 1998b) and in fact the relationship of the Semantic Web to databases parallels the relationship of the World Wide Web to documents *(Fig. 2)*.

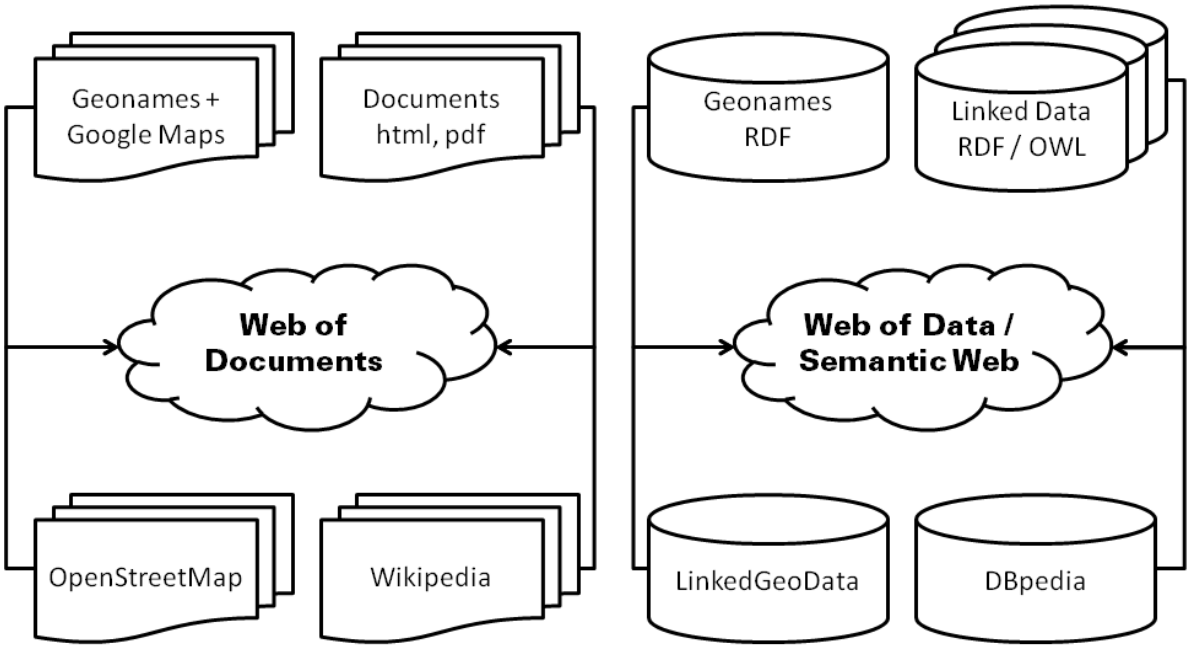

Fig. 2 Relationship between documents and the Web of Documents compared to the relationship of relational databases and the Web of Data / Semantic Web

Hence the Semantic Web can extend the world of databases over the bounds of their domains and furthermore makes the linking of many relating databases possible. Consequently sophisticated queries and operations can be performed across them (Berners-Lee 1998b). This leads to the vision that the interoperable Semantic Web in future can be the "Web 3.0" (Lassila and Hendler 2007).

[6] http://linkeddata.org/

THE SPATIAL SEMANTIC WEB

Though the spatial semantic web is still in its beginnings, projects such as LinkedGeoData (Auer et al. 2007) and applications alike SemaPlorer (*Fig. 3*) (Schenk et al. 2008) and DBpedia mobile (Becker 2008) have already arisen. LinkedGeoData is the RDF representation of the popular OpenStreetMap dataset. SemaPlorer is an application that allows users to visualize large, mixed-quality and semantically heterogeneous geographic data sets. Wikipedia, Flickr and other distributed VGI sources can be explored. DBpedia mobile renders a map and shows nearby locations from the DBpedia dataset.

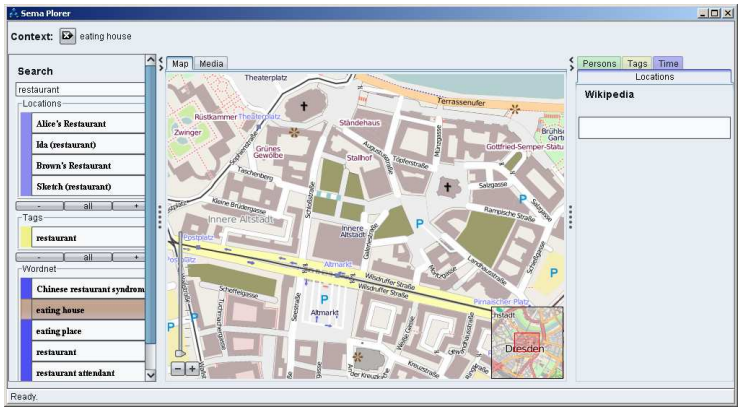

Fig. 3 The Semaplorer Application

Several possible approaches for the integration of spatial data in the Semantic Web have been discussed (Dolbear and Hart 2008, Lieberman and Goad 2008, Auer et al. 2009a), but there is no commonly accepted standard yet. Likewise for tools there are different options. Dolbear and Hart distinguish three different types of tools (Dolbear and Hart 2008):

1. Tools for database to RDF mapping, as implemented in D2RQ (Bizer and Seaborne 2004) and triplify (Auer et al. 2009). All data is stored in relational databases and SQL queries are used to map data to virtual RDF graphs.
2. Semantic Web Services as used in LinkedGeoData (Auer et al. 2009b).
3. RDF triple stores such as Virtuoso (Erling and Mikhailov 2007), Jena (Caroll 2004), Oracle (Lopez and Annamalai 2006), 3store (Harris and Gibbins 2003) and Sesame (Broekstra et al. 2002).

However none of them has gained mainstream usage yet.

For the semantic part of the spatial information it is necessary to find useful ontologies. In the decentralized environment of the web, there is no single world view, but a wide range of possible ontologies. Regarding this it is important to avoid "rebuilding babel with ontologies"[7] *(Fig. 4)*.

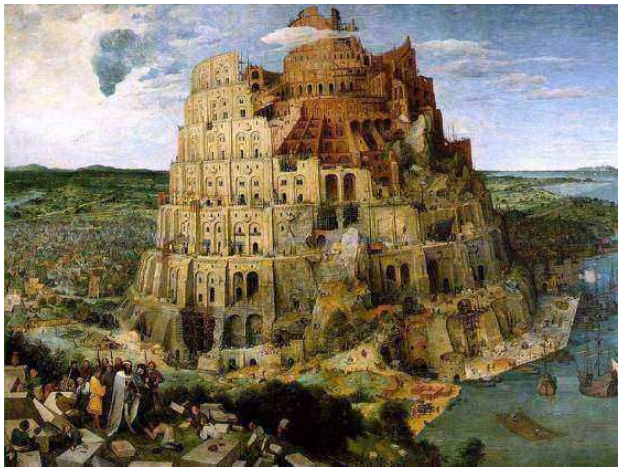

Fig. 4 The Tower of Babel by Pieter Brueghel the Elder

For this reason, research on Semantic reference systems (Kuhn 2003) is important. Their objective is to generate formalized representations of the meaning of geographic features to allow the interoperability between different domains which use shared vocabularies. Research activities focus on finding minimal subsets of elements that define geographic features to which everyone can agree and the development of domain specific profiles, which include more specific descriptions of geographic features.

A benefit of the integration of spatial information into the semantic web is the growing motivation to align different types of spatial and semantic modeling with each other. Furthermore the application of a descriptive logic language like OWL will allow inferences (van Harmelen 2008) within multiple representation datasets.

COMPARISON OF MRDB AND LINKED DATA

In the previous sections we have introduced Linked Data. To discuss the question, whether Linked Data can be a Multiple Representation Database (MRDB) at web scale we will compare the concepts of MRDB and Linked Data with each other. For this purpose we have chosen the definition of MRDB given by Sarjakoski (2007). As *Tab. 1*

[7] http://www.geospatialmeaning.eu/category/semantic-web/

shows, there are basic similarities between both concepts especially, if the definition of MRDB is extended to not only imply a database but also the web to as an underlying environment. Both, MRDB and Linked Data, include different views on the same real world objects and for both a geometry-driven feature matching is applicable to incorporate new data.

These two concepts differ in their focus on level of detail and in semantic as well as geometric abstraction. As the purpose of an MRDB is often high quality and effective map production, it needs to have different levels of abstraction, whereas the purpose of Linked Data initially is not map production, but the pure access to arbitrary spatial and non-spatial information in the web.

	MRDB	**Linked Data**
Similarities	• a (**database | web**) structure in which several representations of the same geographic entity or phenomenon, such as a building or a lake, are stored as different objects in a (**database | web**) environment and linked (Sarjakoski 2007) • consist of various representations […], providing a set of different views of the same object (Sarjakoski 2007) • geometry driven feature matching • matching of database schemas, RDF vocabularies, OWL ontologies	
Differences	• focus on different geometric and semantic abstraction levels • Level of Detail strongly considered • persistence and consistency can supervised by the producer • corresponding objects at different levels are explicitly linked (Sarjakoski 2007) • manuals that contain verbal descriptions of attributes and class hierarchies • corporate data • authority-driven	• focus on different representations of the same entity: different (media) type and content of information • Level of Detail sparsely considered • persistence and consistency cannot be guaranteed by web links • marginal vertical structure of geographic data • use of RDF standard for meta data • web / distributed data • community-driven

Tab. 1 Similarities and differences in the concepts of MRDB and Linked Data

As Linked Data contains data, which is distributed over the web, it is a more community-driven approach then corporate MRDBs, which are mostly maintained by an authority. *Tab. 2* summarizes the comparison and gives and outlook towards upcoming research areas

	MRDB	**Linked Data**
Major Methods	• generalisation	• semantic web technology (RDF, OWL)
Purpose	• high quality and effective map production • derive different type of maps from the representation levels	• access to spatial and non-spatial information • cross domain data access (SPARQL)
Research areas	• automated Generalisation • updates • context Modelling	• formalized ontological descriptions of geographic features (shared vocabularies) • semantic Interoperability • self validating data

Tab. 2 Methods, Purpose and Research areas for MRDB and Linked Data

WORK IN PROGRESS

Ongoing work focuses on the linking of the data sources LinkedGeoData and Geonames. The benefit of this linking will be that the multilingual place gazetter of the Geonames project will enrich the mostly monolingual tagged points of interests within the OpenStreetMap data set. A second advantage will be the possibility of data validation using two independent data sets.

The Geonames dataset contains an explicitly tagged structure of the administrative hierarchy of place names. In the OpenSteetMap dataset the same information is contained implicitly, because of the given geographic extent of the same features. The percentage of places, which fulfill the condition to have the same administrative hierarchy derived from both data sets, will be examined for the purpose of data validation.

Furthermore research is ongoing on using SPARQL query language. Results from querying spatial and non-spatial information contained in the Semantic Web could

serve as input for thematic mapping. Work on tools to publish geographical data as Linked Data in the Semantic Web will support this.

REFERENCES

Auer, S., C. Bizer, and K. Idehen (2009a), The DBpedia Data Set, http://wiki.dbpedia.org/Datasets.

Auer, S., C. Bizer, C. Müller, and A. V. Zhdanova (Eds.) (2007), *Proceedings of the 1st Conference on Social Semantic Web (CSSW)*, Leipzig.

Auer, S., S. Dietzold, J. Lehmann, S. Hellmann, and D. Aumueller, Triplify – Light-Weight Linked Data Publication from Relational databases, in *WWW '09: Proceedings of the 18th international conference on World Wide Web, Madrid, Spain,* pp. 621–630, 2009.

Auer, S., J. Lehmann, and S. Hellmann (2009b), *LinkedGeoData - Adding a Spatial Dimension to the Web of Data*, Universität Leipzig, Institute of Computer Science, Leipzig.

Becker, C. (2008), A Location-Aware Mobile Client for the Semantic Web, *Diplomarbeit*, Freie Universität, Berlin.

Beeri, C., D. Yerach, Y. Kanza, E. Safra, and Y. Sagiv (2005), Finding corresponding objects when integrating several geo-spatial datasets, in *Proceedings of the 13th annual ACM international workshop on Geographic information systems. Session: Data Integration and Data Mining,* pp. 87–96, Bremen, Germany, 2005.

Berners-Lee, T. (1998a), Semantic Web Road Map, http://www.w3.org/DesignIssues/Semantic.html.

Berners-Lee, T. (1998b), What the Semantic Web can represent, http://www.w3.org/DesignIssues/RDFnot.html.

Berners-Lee, T. (1999), Web Architecture: Describing and Exchanging Data, http://www.w3.org/1999/04/WebData.html.

Berners-Lee, T. (2006), Linked Data - Design Issues, http://www.w3.org/DesignIssues/LinkedData.html.

Bizer, C., and A. Seaborne (2004), D2RQ – Treating Non-RDF Databases as Virtual RDF Graphs, in *Proceedings of 3rd International Semantic Web Conference (ISWC04)*, edited by S. A. McIlraith et al., Springer, Hiroshima, Japan.

Brickley, D., and R. V. Guha (2004), RDF Vocabulary Description Language 1.0: RDF Schema, http://www.w3.org/TR/rdf-schema/.

Broekstra, J., A. Kampman, and F. van Harmelen (2002), Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema, in *The Semantic Web — ISWC 2002, Lecture Notes in Computer Science*. 1. ed., pp. 54–68, Springer, Berlin / Heidelberg.

Budhathoki, N. R., B. Bruce, and Z. Nedovic-Budic (2008), Reconceptualizing the role of the user of spatial data infrastructure, *GeoJournal 72,* 149–160, http://www.gsdi.org/gsdi11/papers/pdf/279.pdf.

Caroll, J. J. [. P. L. B. (2004), Jena: Implementing the Semantic Web Recommendations.

Cyganiak, R., and A. Jentzsch (2010), About the Linking Open Data dataset cloud, http://richard.cyganiak.de/2007/10/lod/.

Dolbear, C., and G. Hart (2008), Opportunities in Exploiting Semantics as an Aid to Information Integration, A National Mapping Agency Perspective, in *Creating Spatial Information Infrastructures. Towards the Spatial Semantic Web*, edited by P. van Oosterom and S. Zlatanova, pp. 89–101, Taylor & Franics Group, Boca Raton.

Erling, O., and I. Mikhailov (2007), RDF Support in the Virtuoso DBMS, in *Proceedings of the 1st Conference on Social Semantic Web (CSSW)*, edited by S. Auer et al., pp. 59–68, Leipzig.

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

Harris, S., and N. Gibbins (2003), 3store: Efficient Bulk RDF Storage, in *Proceedings of the 1st International Workshop on Practical and Scalable Semantic Systems (PSSS'03),* pp. 1–20.

Kuhn, W. (2003), Semantic reference systems, *International Journal of Geographic Information Science 17*(5), 405–409.

Lassila, O., and J. Hendler (2007), Embracing "Web 3.0", *IEEE Internet Computing 11*(3), 90–93.

Lassila, O., and R. R. Swick (1999), Resource Description Framework (RDF) Model and Syntax Specification, http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/.

Lieberman, J., and C. Goad (2008), Geosemantic Web Standards for the Spatial Information Infrastructure, Nice to have or hopeless without?, in *Creating Spatial Information Infrastructures. Towards the Spatial Semantic Web*, edited by P. van Oosterom and S. Zlatanova, pp. 119–128, Taylor & Franics Group, Boca Raton.

Lopez, X., and M. Annamalai (2006), Developing Semantic Web Applications using the Oracle Database 10g RDF Data Model, http://www.oracle.com/technology/tech/semantic_technologies/pdf/oow2006_semantics_061128.pdf.

O'Reilly, T. (2005), What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software, http://www.oreillynet.com/lpt/a/6228.

Samal, A., S. Seth, and K. Cueto (2004), A feature-based approach to conflation of geospatial sources, *International Journal of Geographic Information Science 18*(5), 459–489, http://dx.doi.org/10.1080/13658810410001658076.

Sarjakoski, L. T. (2007), Conceptual Models of Generalisation and Multiple Representation, in *Generalisation of geographic information: cartographic modelling and applications*, edited by International Cartographic Association, pp. 11–36.

Schenk, S., C. Saathoff, A. Baumesberger, F. Jochum, A. Kleinen, S. Staab, and A. Scherp (2008), SemaPlorer Interactive Semantic Exploration of Data and Media based on a Federated Cloud Infrastructure, in *7th International Semantic Web Conference ISWC*.

van Harmelen, F. (2008), Semantic Web Technologies as the Foundation for the Information Infrastructure, in *Creating Spatial Information Infrastructures. Towards the Spatial Semantic Web*, edited by P. van Oosterom and S. Zlatanova, pp. 37–52, Taylor & Franics Group, Boca Raton.

von Gösseln, G., and M. Sester (2004), Integration of geoscientific data sets and the german digital map using a matching approach, in *International Archives of Photogrammetry and Remote Sensing,* Istanbul.

Wikipedia (2010), Serendipity - Wikipedia, the free encyclopedia, http://en.wikipedia.org/wiki/Serendipity.