“80% of All Information is Geospatially Referenced”??? Towards a Research Framework: Using the Semantic Web for (In)Validating this Famous Geo Assertion

Stefan Hahmann[1], Dirk Burghardt[1] and Beatrix Weber[2]

[1]Institute for Cartography
[2]Institute for English and American Studies
Dresden University of Technology, Helmholtzstraße 10, D-01062 Dresden
Email: {Stefan.Hahmann, Dirk.Burghardt, Beatrix.Weber}@tu-dresden.de

1. BACKGROUND AND OBJECTIVES

The assertion that 80% of all information has some geographic reference is quite famous amongst geoscientists. It has been repeatedly stated in numerous publications with plenty of modifications, e.g.: “eighty to ninety percent of all the information collected and used was related to geography” (Huxhold 1991: 22–23), “as much as 80% of all information held by business and government may be geographically referenced” (Franklin 1992: 12), “research shows that approximately 80% of all decisions in the public sector are based on georeferenced data” (Riecken 2001: 218) and “95% is more accurate today, new technology is partially responsible, including cell phones, GPS devices and electronic toll collectors” (Perkins 2010).

Authors generally do not mention the methodology of the statistics and refer either to previous academic or to other non-academic sources. Franklin cites an annual report of the Ohio Geographically Referenced Information Program (OGRIP 1990) and Huxhold mentions a brochure of the Municipality of Burnaby, Canada (Municipality of Burnaby 1986). The only study that reports a method is Sussmann (1993). Here entities of all databases of the Canadian City of Scarborough were counted and it was found that 28% of all entities were stored within the 4 GIS databases of the in total 11 databases. Nevertheless even this publication concludes that “the municipal data model in Scarborough demonstrated that over 80% of all data could be associated with geography” without justifying how this number was generated.

In a blog comment, Bob Gaspirc, who was a member of the Toronto Municipal Atlas Group in 1983, claims himself to be one of the original sources of this quote and mentions that it was basically invented for cost justification of IT hardware (Gaspirc 2010).

Today the thesis can still be read in research contexts – “According to the generally accepted assertion that 80% of all information has a reference to space” (Fitzke and Greve 2010: 735) – and hence has become a kind of truism in geographic information science.

The objective of this paper is to introduce a research framework that aims to prove or disprove this famous assertion. For this purpose we will first look at the definition of the terms *spatial* and *geospatial reference* from a state-of-the-art perspective. Section 3 describes the foundations of the Semantic Web, which will be used as a tool for validation or invalidation. Section 4 presents the actual approach and shows figures that visualise it. This is followed by a preliminary results section based on a small dataset to exemplify the approach. In the last section we present detailed research questions for the topic, which need to be addressed in our future work.

2. DEFINITION OF THE TERMS SPATIAL AND GEOSPATIAL REFERENCE

A German textbook for cartographers defines the term *spatially referenced information* (“raumbezogene Information”) as every piece of information that provides a geometric assignment for an object within a reference system (Hake et al. 2002: 4). Here spatial reference is defined in very general terms. Bollmann refines this and defines the term *spatial reference* in the context of cartography as a property of objects concerning their *geospatial* relation (Bollmann 2002: 266). It can be seen that Hake’s definition puts more weight on *space* in general whereas Bollmann adds *geo* to *spatial* as this is the context where geoscientists usually use the term. The threshold that separates spatial and geospatial information in these definitions is a matter of scale.

Furthermore Bill (2008) distinguishes *direct* and *indirect* spatial reference. Direct spatial reference is established by two or three dimensional coordinates within a spatial reference system. Indirect spatial reference is fuzzier and less accurate. It can be described by e.g. associated postcodes, addresses, geographic names or administrative areas, which themselves can be referenced with direct spatial information.

In our research work we will concentrate on *geospatial reference* as defined by Bollmann. Furthermore we will use Bill's classification of direct and indirect geospatial reference. We will add the categories *geospatial coordinates* and *non-geospatial information*, i.e. information without geospatial reference.

3. THE SEMANTIC WEB

Semantic Networks have been applied by researchers of artificial intelligence and information science (Sowa 1987). They describe and graphically represent concepts and the relations between them.

Berners-Lee et al. (2001) introduced the idea of the *Semantic Web*. This is an implementation of a Semantic Network that uses methods of the World Wide Web, such as URIs to describe concepts or things and hyperlinks that define relations between them. The inventors of the Semantic Web aim to advance the World Wide Web that was initially made for human consumption from a "machine-readable" to a "machine-understandable" web (Lassila and Swick 1999).

Within the Semantic Web the Resource Description Framework (RDF) can be used to describe data semantically (Brickley and Guha 2004). In RDF, every triplet of information consists of three resources – a subject, an object and a predicate that describes the relationship between them. Figure 1 shows an example taken from the LinkedGeoData project. As an exception an object may not be a resource, but also a literal. RDF allows data providers to publish their data and to link it to other data, including data of different knowledge domains. Figure 2 illustrates how information of Geonames, OpenStreetMap and Wikipedia, or more precisely information of the Semantic Web counterparts of these projects – Geonames RDF, DBpedia and LinkedGeoData – are linked with each other. Links between information entities are RDF predicates that indicate the relationship between the two connected items.

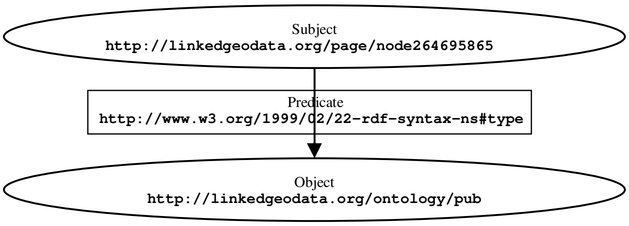
Subject
http://linkedgeodata.org/page/node264695865
Predicate
http://www.w3.org/1999/02/22-rdf-syntax-ns#type
Object
http://linkedgeodata.org/ontology/pub

Figure 1: An example RDF triplet showing that the LinkedGeoData node 264695865 is of the RDF type LinkedGeoData pub.

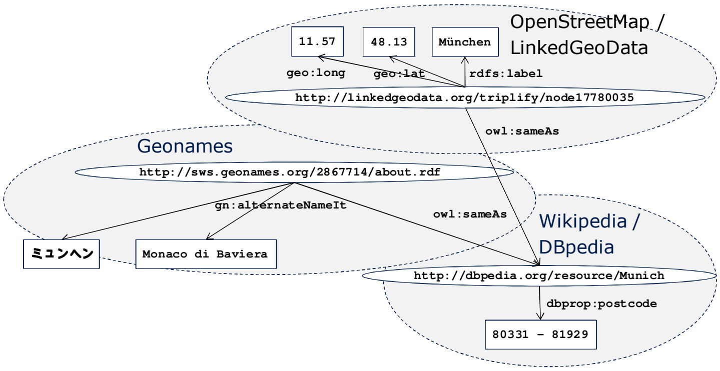

Figure 2: An RDF graph showing linked data from different domains.

Once RDF data has been published on the web and linked to other data, this data is called “Linked Data” (Berners-Lee 2006). The Linked Open Data[1] project aims to provide pointers to at least all openly accessible linked data. Figure 3 shows an illustration of the datasets, which are part of the Linked Open Data cloud, and their cross-references. We tried to assign every dataset to one general field of knowledge. It can be seen that a significant portion of the Linked Open Data cloud is geo related. A big interlinking hub within the Semantic Web is the project DBpedia (Bizer et al. 2009) that provides information from Wikipedia. The biggest geo dataset is the translation of OpenStreetMap for the Semantic Web – LinkedGeoData (Auer et al. 2009).

Hahmann and Burghardt (2010) show similarities and differences of Linked Data and geo databases. An important difference, which is not mentioned in this publication, is that Linked Data primarily aims to formally represent knowledge of different domains whereas geo databases are more suitable for map production in GIS analysis. Another difference concerns data structure. Geo databases generally use relational data models, whereas the Semantic Web is based on a graph data structure. This is important for the proposed method in section 4.

To measure geospatial reference by using the provided linked data, it is important to know how geospatial coordinates can be identified within the linked data cloud. The meaning of information within the Semantic Web is defined by vocabularies. Currently the Semantic Web community uses two different vocabularies, the W3C Basic Geo Vocabulary (W3C Semantic Web Interest Group 2006) and the W3C Geospatial Vocabulary (Liebermann et al. 2007). The W3C Basic Geo Vocabulary is used by the projects DBpedia and LinkedGeoData. It can describe point coordinates using latitude and longitude values within the WGS84 reference system. The W3C Geospatial vocabulary is an extension that can also express lines, polygons and featuretypes. However both vocabularies, whose inventors had initially intended to make them W3C recommendations, have not become official W3C standards yet.

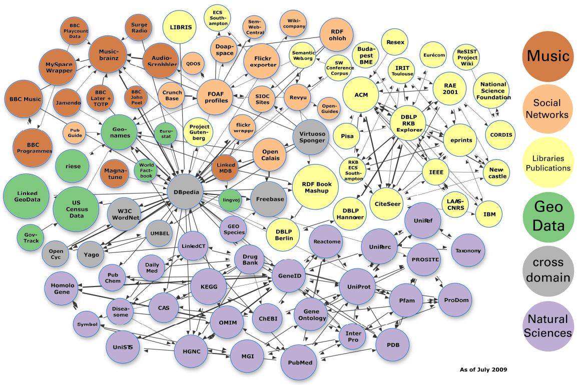

Figure 3: The Linked Open Data cloud. Each color indicates a different field of knowledge. Modified after Cyganiak and Jentzsch (2010).

Furthermore there are two ISO standards: Dublin Core Metadata (ISO 15836) and Geographic Information - Metadata (ISO 19115). Both could also be used to express geospatial coordinates. Nogueras-Iso et al. (2004) present a comparison of these two standards. However both of them have not been adopted by the Semantic Web Community.

4. AN APPROACH FOR MEASURING GEOSPATIAL REFERENCES OF INFORMATION

Kuhn (2010) comments on the 80% geo assertion that “the question itself, […], is very hard to ask in a testable way – what exactly do you count, and what constitutes a spatial reference?” Morley (2010) adds that “the scope of ‘all data’ would be a difficult one to assess – such research might start with some limited types of databases.”

These two statements are also essential from our point of view. As it is not possible to analyse ‘all data’, a corpus of information which can appropriately represent the scope of all data is needed to simplify the problem. We propose to use the Semantic Web for this purpose. To our point of view it meets with all the three mentioned requirements:

[1] http://linkeddata.org/

1. It may prove to be representative for all data, as it includes information from many different fields of knowledge.
2. It allows identifying geospatial coordinates by querying for information that is annotated with terms from one of the mentioned geo vocabularies.
3. It explicitly models references to other information and hence it allows identifying information that is geospatially referenced.

More specifically we propose to use information from the Linked Open Data cloud, which is part of the Semantic Web. This data has the advantage that it is freely accessible for such research. As the Semantic Web is a network of data, we expect that eventually any information will be linked somehow to geospatial coordinates. This hypothesis follows the very similar phenomenon of the six degrees of separation in social networks described by Milgram (1967). In analogy to this we introduce *degrees of geospatial reference*.

Figure 4 illustrates the approach. It shows different datasets at different degrees of geospatial reference. While the degree of geospatial reference is increased, the amount of datasets that have a higher degree of geospatial reference increases likewise. Different colors of the cross-references between the datasets indicate different degrees of geospatial reference. At higher degrees of geospatial reference datasets that have a low degree of geospatial reference are shown in conjunction with datasets that have a higher degree of geospatial reference.

For this visualisation, databases and their cross-references from the Linked Open Data cloud (Cyganiak and Jentzsch 2010) were used. We tried to categorise the datasets on the level of the whole database. This is an extremely simplified model and cannot reveal robust results though it is suitable to illustrate the approach on an acceptable level of detail. Our future research, which aims to find robust results, needs to run the analyses on the level of single information entities.

The degree of geospatial reference of a single information entity or – as in our example – a whole dataset equals the number of edges that are on the shortest path between that information entity or dataset and any geospatial coordinate or geospatial dataset. In order to calculate different degrees of geospatial reference, we applied Dijkstra’s algorithm (Dijkstra 1959) to the Linked Open Data graph. This algorithm is appropriate, as it is capable of providing exact solutions for shortest path problems. Such network analyses can be performed with several GIS software products or even with graph databases that implement this algorithm. By this method it is possible to derive a formal measurement for the degree of geospatial reference of information.

In order to use this measurement to delineate the categories ‘direct geospatial information’, ‘indirect geospatial’ information and ‘non-geospatial information’ an empirical study is required. Our idea is that test persons will be asked to assign selected information entities from the Semantic Web to the three proposed categories of geospatial reference. Subsequently it is possible to calculate the mean path distance within the Semantic Web graph for all three categories. The two thresholds that delineate the categories can be determined by the arithmetic mean of the path distances of the two adjacent categories. If it is possible to find proper thresholds by this method, then it will also be possible to compute ratios of the amount of information of each geospatial reference category in relation to all examined information and hence to validate or invalidate the initial geo assertion.

5. PRELIMINARY RESULTS

In this section, we present the results of our preliminary study, which analysed the Linked Open Data cloud on the level of categorised databases. Our aim in presenting the following equations and results is not to reveal real findings but to present the idea of what formal measures need to be computed in future research, which will employ not only a demonstration dataset. At first we used equation 1 to calculate the network average clustering coefficient of the Linked Open Data cloud, which describes the degree to which nodes of the graph are interlinked. The results are shown in Table 1.

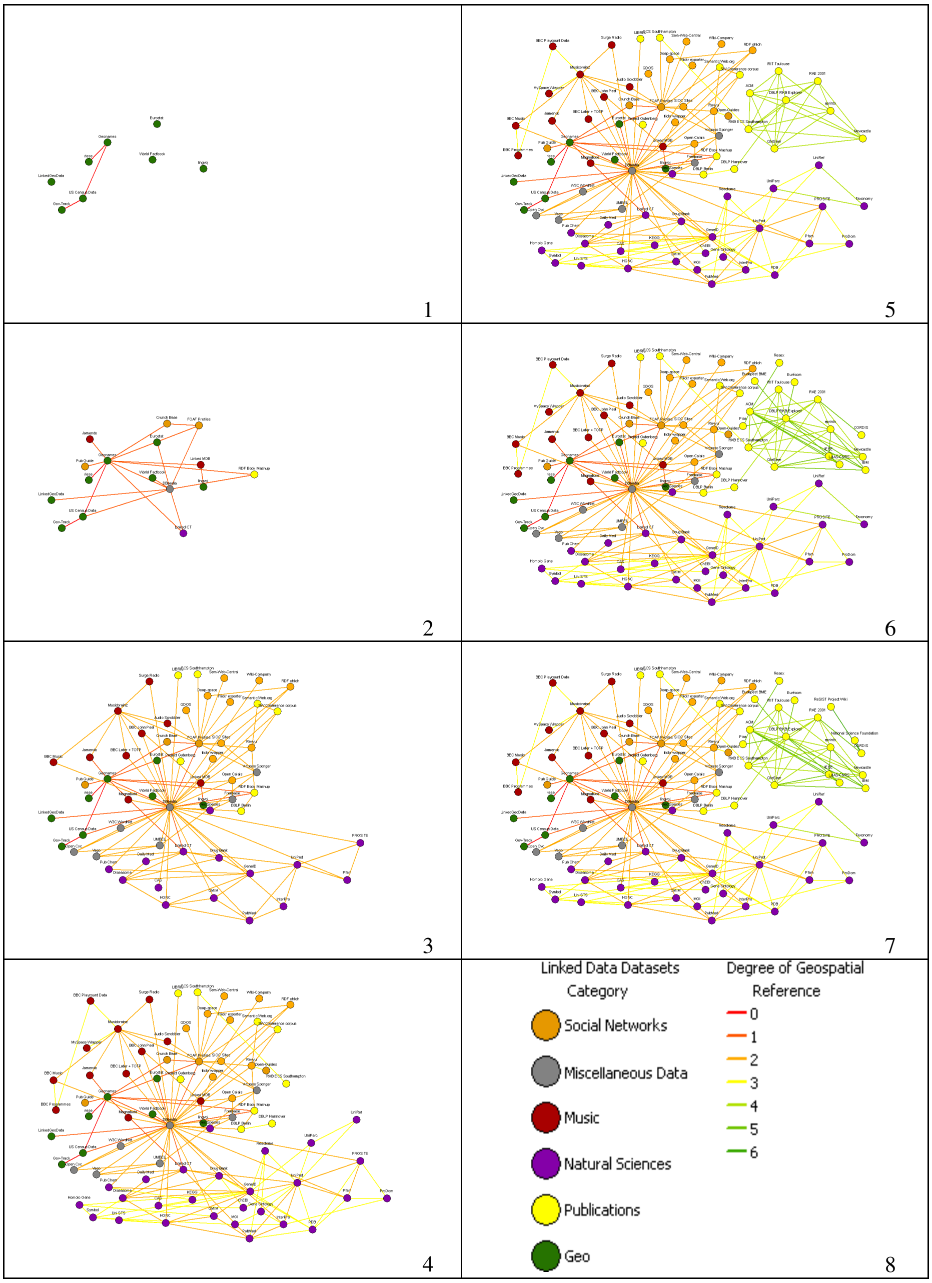

Figure 4: Visualisation of the approach. Datasets of the Linked Open Data cloud and its cross-references. An increasing *degree of geospatial reference* results in an increasing portion of *geospatially referenced information*.

$$network\ average\ clustering\ coefficient = \frac{2 \times \sum edges}{\sum nodes \times (\sum nodes - 1)} \quad (1)$$

Nodes	95
potential edges	4495
existing edges	195
network global average clustering coefficient (%)	4

Table 1: Calculation of network average clustering coefficient.

Despite this rather small degree of linkage, Table 2 shows that only six degrees of geospatial reference can be identified in our categorised Linked Open Data cloud. For each degree of geospatial reference we calculated the ratio of affected nodes and edges to all nodes and edges of the graph of the Linked Open Data cloud. Equations 2 and 3 were used for this calculation.

$$ratio_i = \frac{\sum nodes_i}{\sum nodes_{i=6}} \quad (2) \qquad ratio_i = \frac{\sum edges_i}{\sum edges_{i=6}} \quad (3)$$

degree of geospatial reference i	0	1	2	3	4	5	6
Nodes	8	16	59	77	85	93	95
Ratio$_{Nodes}$	8	17	62	81	89	98	100
Edges	3	20	107	150	177	193	195
Ratio$_{Edges}$	2	10	55	77	91	99	100

Table 2: Degrees of geospatial reference and ratios of affected nodes and edges.

As described in section 4 the task of further research will be to find the thresholds between the different numeric degrees of geospatial reference that delineate the verbal categories of geospatial reference. The results of Table 2, especially the ratio of nodes affected by the different numeric degrees of geospatial reference, can then be taken to answer the initial question of how much information is geospatially referenced.

The last result is shown in Table 3. This is the average path distance of any information entity to a geospatial information entity within the Semantic Web. This equals the average degree of geospatial reference. This measurement is not primarily important to answer the initial research question but is a statistical value that can be interpreted as: ‘In the Semantic Web the average information entity is no more than three steps away from a geospatial information entity.’ The Calculation was done using equation 4.

$$average\ degree\ of\ \ geospatial\ reference = \frac{\sum_{i=0}^{i=6} (\sum nodes_i - \sum nodes_{i-1}) \times i}{\sum nodes_{i=6}} \quad (4)$$

average degree of geospatial reference	2,4

Table 3: Resulting average degree of geospatial reference within the Linked Open Data cloud.

6. RESEARCH QUESTIONS

Our proposed method uses Dijkstra’s algorithm and the Semantic Web to establish a way of formally measuring the degree of geospatial reference of information. For the purpose of exemplifying this method, we have calculated preliminary results using a demonstration dataset. As discussed in section 4, our future research needs to apply this method on the level of single information entities instead of applying it to whole datasets, as whole datasets can contain many different types of information entities that themselves can have different degrees of geospatial reference. Furthermore, empirical research is needed to link the formal measurement of geospatial reference to how humans perceive the categories ‘direct geospatial reference’, ‘indirect geospatial reference’ and ‘non-geospatial reference’.

For our future work we derive eight research questions, which can be grouped into three sections:

a) Theoretical background for the definition of the term geospatial reference by using the Semantic Web:
   1. To what extent is the Semantic Web applicable as a tool to determine the ratio of geospatially referenced information in relation to all information?
   2. To what extent is it possible to unambiguously identify geospatial coordinates with the vocabularies of the Semantic Web?
   3. Which method is appropriate to reduce the corpus of Linked Data to a manageable amount of data for the study?
   4. How do different predicates that link resources of the Semantic Web influence the result?

b) Empirical and analytical studies to define the term geospatial reference by using the Semantic Web:
   5. Which degree of geospatial reference can be determined for arbitrary resources of the Semantic Web?
   6. Which thresholds can be identified to distinguish information with direct and indirect geospatial reference and non-geospatial information?
   7. How many percent of all information has a geospatial reference?

c) Applicable methods of the Semantic Web:
   8. To what extent can SPARQL, the query language of the Semantic Web, be used to support answering the research questions of section b)?

Research on this topic might raise another – though not that scientific – question:

   9. Should our results be published, if it turns out, that less than 80% of all information has a geospatial reference?

John B. Fagan, Microsoft program manager for Bing Maps and MultiMap products, announced his very clear opinion to this question on Twitter (Fagan 2010): “that geo quote keeps us all in our jobs. Best not go poking around to see if it’s true.”

BIBLIOGRAPHY

Auer, S., J. Lehmann, and S. Hellmann (2009), LinkedGeoData - Adding a Spatial Dimension to the Web of Data, in *Proceedings of the 8th International Semantic Web Conference,* Fairfax, Virginia, USA, pp. 731-746.

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

Berners-Lee, T., J. Hendler, and O. Lassila (2001), The Semantic Web, *Scientific American*, 284(5), http://www-sop.inria.fr/acacia/cours/essi2006/Scientific American_ Feature Article_ The Semantic Web_ May 2001.pdf.

Bill, R. (2008), Raumbezug (Spatial Reference), http://www.geoinformatik.uni-rostock.de/einzel.asp?ID=1004513274.

Bizer, C., J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann (2009), DBpedia - A crystallization point for theWeb of Data, *Journal of Web Semantics*, 7(3), pp. 154-165.

Bollmann, J. (2002), Raumbezug, in *Lexikon der Kartographie und Geomatik*. 1. ed., edited by J. Bollmann and W. G. Koch, Spektrum Akademischer Verlag, Heidelberg, p. 266.

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

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

Dijkstra, E. W. (1959), A note on two problems in connexion with graphs, in *Numerische Mathematik 1,* pp. 269–271.

Fagan, J. B. (2010), Twitter Status, http://twitter.com/johnbfagan/status/9484856028.

Fitzke, J., and K. Greve (2010), Frei oder umsonst? - Nutzergenerierte Geoinformation zwischen Freiheit und Kostenlosigkeit, in *Angewandte Geoinformatik - 22. AGIT-Symposium*. 1. ed., Wichmann, Berlin, pp. 732–741.

Franklin, C. (1992), An Introduction To Geopgrahic Information Systems: Linking Maps To Databases, *Database: the magazine of database reference and review*, 15(2), pp. 10-22.

Gaspirc, B. (2010), Comment on Blogpost "The source of the assertion that 80% of all organisational information is geographic", http://povesham.wordpress.com/2010/02/22/the-source-of-the-assertion-that-80-of-all-organisational-information-is-geographic/#comment-4487.

Hahmann, S., and D. Burghardt (2010), Connecting LinkedGeoData and Geonames in the Spatial Semantic Web, in *Proceedings of extented abstracts,* Zurich, Switzerland.

Hake, G., D. Grünreich, and L. Meng (2002), *Kartographie. Visualisierung raum-zeitlicher Informationen*, Walter de Gruyter & Co, Berlin.

Huxhold, W. E. (1991), *An Introduction to Urban Geographic Information Systems*, Oxford University Press, New York.

International Organization for Standardization (ISO) (2003), Geographic Information - Metadata *19115:2003.*

International Organization for Standardization (ISO) (2009), Information and documentation - The Dublin Core metadata element set *15836:2009.*

Kuhn, W. (2010), Comment on Blogpost "The source of the assertion that 80% of all organisational information is geographic", http://povesham.wordpress.com/2010/02/22/the-source-of-the-assertion-that-80-of-all-organisational-information-is-geographic/#comment-2141.

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/.

Liebermann, J., R. Singh, and C. Goad (2007), W3C Geospatial Vocabulary, Final Report, http://www.w3.org/2005/Incubator/geo/XGR-geo/.

Milgram, S. (1967), The small world problem, *Psychology Today*, 1(2), pp. 60-70.

Morley, J. (2010), Comment on Blogpost "The source of the assertion that 80% of all organisational information is geographic", http://povesham.wordpress.com/2010/02/22/the-source-of-the-assertion-that-80-of-all-organisational-information-is-geographic/#comment-2133.

Municipality of Burnaby (1986), Invitation to Information, Brochure, Burnaby, Canada.

Nogueras-Iso, J., F. J. Zarazaga-Soria, J. Lacasta, R. Béjar, and P. R. Muro-Medrano (2004), Metadata standard interoperability: application in the geographic information domain, *Computers, Environment and Urban Systems*, 28(6), pp. 611-634.

OGRIP (1990), Advisory Committee's First Year Report, Department of Administrative Services, Columbus, OH.

Perkins, B. (2010), Have you mapped your data today? http://www.computerworld.com/s/article/350588/Have_You_Mapped_Your_Data_Today_.

Riecken, J. (2001), The improvement of the access to public geospatial data of cadastral and surveying and mapping as a part of the development of a NSDI in Northrhine-Westfalia, Germany, in *Proceedings of the 4th AGILE Conference on GIScience*, AGILE, Brno pp. 215–221.

Sowa, J. F. (1987), Semantic Networks, in *Encyclopedia of Artificial Intelligence*. 2. ed., Wiley, New York.

Sussmann, R. (1993), Municipal GIS and the enterprise data model, *International Journal of Geographical Information Science*, 7(4), pp. 367-377.

W3C Semantic Web Interest Group (2006), Basic Geo (WGS84 lat/long) Vocabulary, http://www.w3.org/2003/01/geo/.