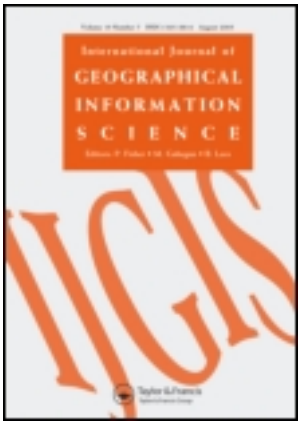

International Journal of Geographical Information Science

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tgis20

How much information is geospatially referenced? Networks and cognition

Stefan Hahmann [a] & Dirk Burghardt [a]

[a] Institute for Cartography , Dresden University of Technology , Dresden, Germany

Published online: 23 Nov 2012.

To cite this article: Stefan Hahmann & Dirk Burghardt (2013): How much information is geospatially referenced? Networks and cognition, International Journal of Geographical Information Science, 27:6, 1171-1189

To link to this article: http://dx.doi.org/10.1080/13658816.2012.743664

International Journal of Geographical Information Science, 2013
Vol. 27, No. 6, 1171–1189, http://dx.doi.org/10.1080/13658816.2012.743664

How much information is geospatially referenced? Networks and cognition

Stefan Hahmann* and Dirk Burghardt

Institute for Cartography, Dresden University of Technology, Dresden, Germany

(*Received 27 June 2012; final version received 21 October 2012*)

The aim of this article is to provide a basis in evidence for (or against) the much-quoted assertion that 80% of all information is geospatially referenced. For this purpose, two approaches are presented that are intended to capture the portion of geospatially referenced information in user-generated content: a network approach and a cognitive approach. In the network approach, the German Wikipedia is used as a research corpus. It is considered a network with the articles being nodes and the links being edges. *The Network Degree of Geospatial Reference* (*NDGR*) is introduced as an indicator to measure the network approach. We define NDGR as the shortest path between any Wikipedia article and the closest article within the network that is labeled with coordinates in its headline. An analysis of the German Wikipedia employing this approach shows that 78% of all articles have a coordinate themselves or are directly linked to at least one article that has geospatial coordinates. The cognitive approach is manifested by the *categories of geospatial reference* (*CGR*): direct, indirect, and non-geospatial reference. These are categories that may be distinguished and applied by humans. An empirical study including 380 participants was conducted. The results of both approaches are synthesized with the aim to (1) examine correlations between NDGR and the human conceptualization of geospatial reference and (2) to separate geospatial from non-geospatial information. From the results of this synthesis, it can be concluded that 56–59% of the articles within Wikipedia can be considered to be directly or indirectly geospatially referenced. The article thus describes a method to check the validity of the '80%-assertion' for information corpora that can be modeled using graphs (e.g., the World Wide Web, the Semantic Web, and Wikipedia). For the corpus investigated here (Wikipedia), the '80%-assertion' cannot be confirmed, but would need to be reformulated as a '60%-assertion'.

Keywords: geospatial reference; geographic information retrieval; scale-free networks; cognition of geographic information; Wikipedia

1. Introduction

This article was motivated by the hype about a figure that is as frequently quoted as it is empirically unsupported: '80% of all information contains some geospatial reference'. This 'geo-assertion' that occurs in various guises is widely known among members of geoscience and geo-business communities. Today, it has become so entrenched that it almost passes for fact. Some of the feedback that we have received from the survey conducted in preparation of this article showed us that there is a demand within the GIS community

*Corresponding author. Email: stefan.hahmann@tu-dresden.de

to test this figure by scientific means. Hence, almost 30 years after its first occurrence, it is time to reexamine the assertion, even if such a study only touches upon the margins of classic GIScience research. In any case, such an investigation might still be assigned to the area of 'cognition of geographic information', which has been recognized as one of the sub-fields of GIScience (cf. Mark 2003).

Today, it seems to be impossible to doubtlessly ascertain where the assertion originally occurred. In Hahmann *et al*. (2011), we have reported that the earliest traceable references for this piece of 'geo-folk-wisdom' go back to the 1980s and originated in Canada/United States during the rise of GIS. Huxhold (1991, pp. 22–23) cites the assertion from a brochure of the Municipality of Burnaby, Canada (Municipality of Burnaby 1986) 'as much as 80% of all information held by business and government may be geographically referenced'. We have also collected information on later occurrences of the assertion in scientific and non-scientific publications, for example, 'as much as 80% of all information held by business and government may be geographically referenced' (Franklin 1992, p. 12), 'research shows that approximately 80% of all decisions in the public sector are based on georeferenced data' (Riecken 2001, p. 218), '95% is more accurate today, new technology is partially responsible, including cell phones, GPS devices and electronic toll collectors' (Perkins 2010), and 'According to the generally accepted assertion that 80% of all information has a reference to space . . . ' (Fitzke and Greve 2010, p. 735, translated from German). Numerous further references could easily be listed.

The sources of the 80%-assertion indicate that the figure applied to municipal and government data, meaning that the initiators of the claim mainly referred to coordinates, identifiers, and address data. Since its early occurrences, the assertion has been overgeneralized and many authors have blindly copied it, even though it has no basis in evidence. People who know the assertion express a tension between hoping the number is realistic and doubts as to whether it has been invented to drive the market.

If we further develop the idea Vanessa Lawrence expresses as 'everything happens somewhere' (Lawrence 2009, p. 4), then we may, of course, conclude that 100% would be the correct answer. Likewise, in Hahmann *et al*. (2011), we discussed that this applies almost entirely to information that is stored in a network structure such as the Semantic Web or the World Wide Web (WWW). That is due to the fact that in these types of networks, any piece of information – apart from only a few isolated items – is linked to any other in a giant, single connected component, as Broder *et al.* (2000) have shown for the WWW. Hence, given a network where at least some geospatial information is stored, it will be possible to find a path from any item to some piece of geospatial information. These paths might be regarded as realizations of geospatial reference.

However, by 'geospatially referenced' something more restrictive is meant. Thus, what really matters is to answer the question for which proportion of information the geospatial component is in any way relevant? Extending the network of knowledge analogy to the human mind, we may ask: where does human cognition cut off the paths that link arbitrary information with geospatial information? There may be items that can be said to be geospatially referenced *per se*, such as geographic names. However, a wide range of things, such as, persons, historic events, companies, cultural artifacts (books, music, movies, art, etc.), raw materials, food items, or political views only become geospatially referenced depending on the context where these things are used.

If we want to answer the question of how much information is geospatially referenced, we need to think about both the numerator and the denominator, that is, we have to determine how many percent and of what? For the analytical and empirical studies that we present in this article, we decided to use Wikipedia, as it can be considered a corpus of

domain neutral knowledge. Hecht and Moxley (2009) have used it to evaluate Tobler's first law of geography.

2. Terms

Before we explain our methods in detail, we want to briefly discuss the terms that we use in this article.

(1) *Geospatial*: We consider it important to use the term '*geospatial*' in contrast to just 'spatial', as the 80%-assertion clearly referred to geospatial information. The threshold between these two concepts is a matter of scale. See Haggett (2001, pp. 21–23) for a delineation of the range of scales that is of interest for geographic science and a discussion of 'magnitudes of geographic order'.

(2) *Information*: We are not going to use the word 'information' in the sense introduced by Shannon and Weaver's information theory (Shannon and Weaver 1949). As discussed by Goodchild (2003), this notion of the word 'information' is mainly concerned with the form and coding of messages, which is only of limited use for GIScience. Instead, we use Goodchild *et al.*'s (1999) approach to geographic information, who regard a tuple $<x, y, z, t, U>$[1] as the primitive element of geographical information in space-time. This includes that we do not consider toponyms alone (e.g., 'Mount Everest') to be sufficient identifiers for positions on the Earth's surface, as they need to be converted into a location by a gazetteer first (cf. Goodchild 2003). If geographical information is assessed via this space-time-attribute tuple approach, single tuples, that is, pieces of geographical information, may be counted. Moreover, we prefer using the term 'information' to using the term 'data', as data chiefly refer to numerical information items, which will not be the main focus of our study. As the *z*- and the *t*-components of the mentioned tuple are not relevant for our approach, they will not be considered further within this article.

(3) *Reference/referenced*: There are two meanings of the term 'reference', which are both used in this article: (a) (geospatially) 'referenced' in the sense of having coordinates or being based on exact measurement within a defined (geospatial) reference system and (b) (geospatial) 'reference' in the sense of a general relation to geospace or geography, which allows a potential transformation into coordinates at varying levels of granularity. The first sense is implied in conjunction with the network approach and the latter in conjunction with the cognitive approach. Both approaches will be introduced in the following two sections. In combination with the 80%-figure, meanings (a) and (b) can be found (cf. Hahmann *et al*. 2011).

3. Network approach to geospatial reference

3.1. Related work

Apart from focusing on coordinates, as we do within the network approach, there may be different ways of analytically capturing the portion of geospatially referenced information within a corpus of information. In a previous work, the portion of web documents that contained a US zip code has, for example, been determined (4.5%) by examining a partial web crawl (McCurley 2001). Furthermore, an analysis of geographic entities (toponyms) within newspaper articles yielded the result that, on an average, 75% of the investigated documents contained at least one geographic entity (Cardoso 2011). However, we are not

going to rely on toponyms within the network approach, since we deem the decisions of the Wikipedia authors to assign coordinates to an article as a whole to be a stronger indication of the respective articles being considered geospatial than occurrences of toponyms within the full text of the article.

The network structure of Wikipedia, which is crucial for the network approach, has been investigated by Zlatić *et al.* (2006), Capocci *et al.* (2006), and Voss (2005). It has been shown that the network of Wikipedia articles may be described by the Small-World model developed by Watts and Strogatz (1998), and Zlatić *et al.* (2006) found 4.53 as the average path length between two arbitrary Wikipedia articles by analyzing the 11 largest Wikipedias. Furthermore, they have shown that this number does not significantly differ within different language versions of Wikipedia, despite their different evolutionary stages and – hence – their different sizes. The fact that the main network properties are constant allows us to deduce that the network approach is mainly independent from the evolutionary stage as well as from the language version of the examined Wikipedias, as long as the portion of geospatial articles be approximately equal for each Wikipedia that is analyzed. However, results of Hecht and Moxley (2009) and Dahinden (2011) show that, in fact, this portion varies considerably from language version to language version as well as with regard to different stages of one and the same Wikipedia. We have therefore tested the impact of a lower coverage of geospatial articles on the network approach by a simulation[2] that proved it, however, to be negligible.

3.2. Methodology

In contrast to our suggestion in Hahmann *et al*. (2011), which was to use and analyze the Semantic Web, we have decided to focus on Wikipedia and the 'Wikipedia Article Graph (WAG)' (Hecht and Moxley 2009) in this study. This is for three reasons:

(1) Semantic Web data sources are widespread. Hence, in order to be able to run large-scale analyses with sufficient performance, it would take a significant effort to integrate all these data sources in a local environment. Wikipedia dump files are more convenient: they can easily be downloaded and they can be parsed by existing tools that support extraction of the internal link structure, such as WikAPIdia (Hecht and Gergle 2010). It would have been possible to use DBpedia (Bizer *et al.* 2009), which contains semantically structured information from the Wikipedia page infoboxes. However, DBpedia lacks article full texts and hence links between article contents. We think that these links are at least as important as links between page infobox information items since both types of links may constitute geospatial reference, which makes both of them important for the network approach to geospatial reference.
(2) In a study, where, as we stated above, the 'what' needs to be clearly delineated, it seems to be a bad idea to use sources that are not really well-closed as would be the case for the Semantic Web. As its subject is defined more precisely, Wikipedia differs from the Semantic Web in this regard. By using Wikipedia as a research corpus, we set the limit for what we study more on the content level and less by technical constraints as we would if we used the Semantic Web.
(3) We have dismissed the Semantic Web, because we found it difficult to translate triple entities into something that would be feasible for study participants with hardly any background knowledge about the Semantic Web, without introducing biases by this translation.

Though we have diverged from our initial plans with regard to the subject of our analysis, we have nevertheless adhered to our proposed method of determining the Degree of Geospatial Reference (cf. Hahmann *et al*. 2011) implementing it for the first time. In order to distinguish the results of the network approach from those of the cognitive approach, which will be described later in this article, we use the term *Network Degree of Geospatial Reference* (*NDGR*). NDGR is the indicator to measure the network approach to geospatial reference. We consider the network approach a model that is capable of analytically capturing the 'geospatiality' of information entities. Given a graph with geospatial information nodes and other nodes that are connected directly or indirectly, NDGR can be determined using breadth-first labeling of the network, as all distances are integers. For our study, we assign a value of 0 to the NDGR indicator for all articles that use a coordinate template in their headlines. Figure 1 shows an example of such an article from the German Wikipedia. We also call these articles *geospatial articles*.

Values greater than 0 are assigned under the following condition: an article is assigned a value of 1, if there is at least one link within the full text of this article that connects to an article with NDGR 0. NDGR 2 is constituted if an article has at least one link pointing to an article with NDGR 1. A link in the opposite direction – from an NDGR 0 article to another article – is not sufficient to make that second article NDGR 1. Though these links may also be considered relevant, we do not examine them in this work for reasons of simplification. Furthermore, we do not use weights, neither based on link counts nor on link semantics. So the graph model, we have used for the network approach calculations, can be described as a *directed unweighted graph*. Figure 2 illustrates the network approach.

As the average distance between two arbitrary articles within a Wikipedia is 4.53 (as stated above), we expect the average NDGR to be clearly less than this value, as we do not seek for all possible connections between two nodes but only for those that connect non-geospatial articles to their closest geospatial articles.

Figure 1. Article from the German Wikipedia that uses a coordinate template in its headline.

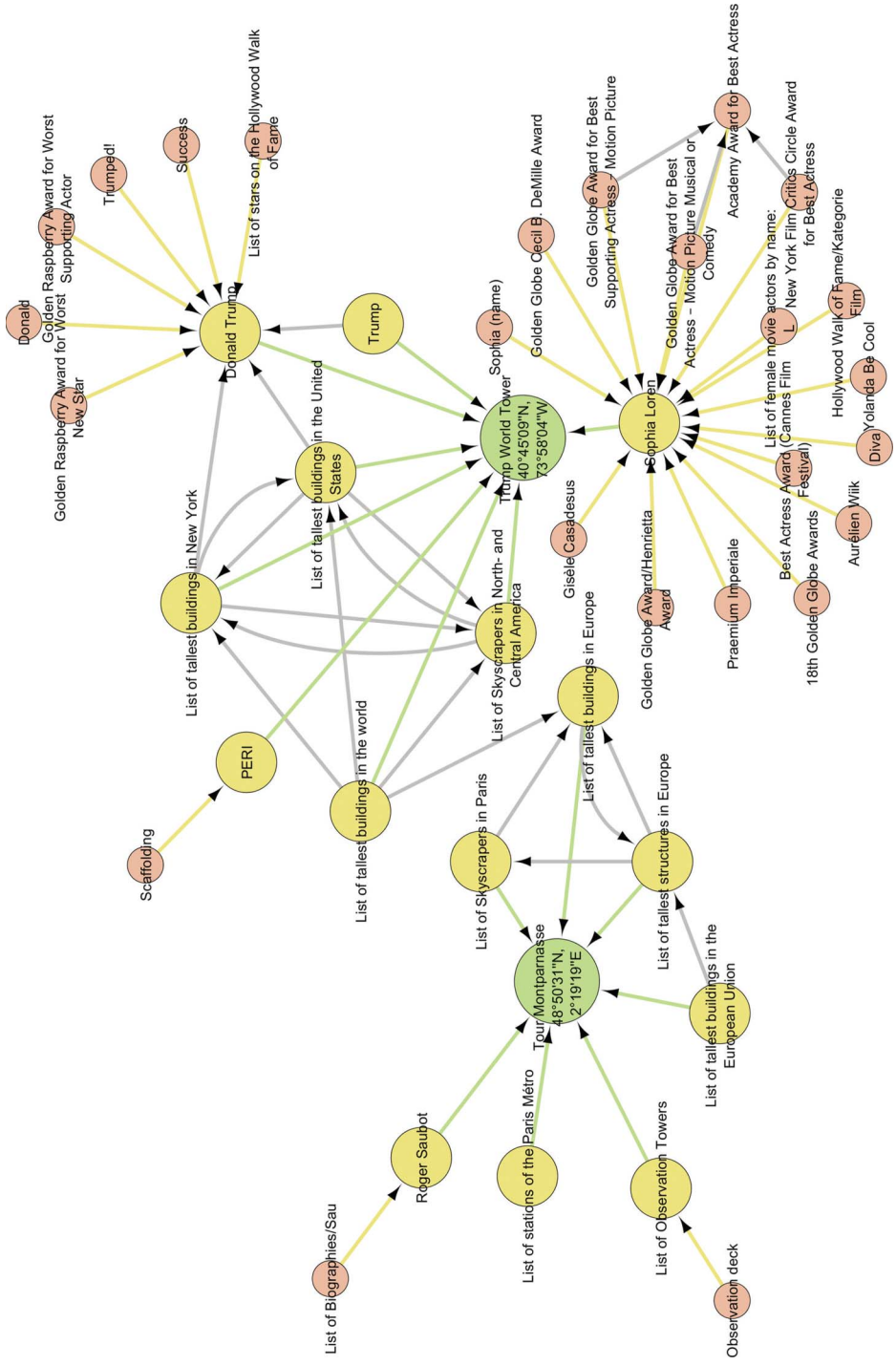

Figure 2. Articles (circles) with NDGR = 0 (green), NDGR = 1 (yellow), and NDGR = 2 (red). Green arrows show references that constitute NDGR 0, yellow arrows show references that constitute NDGR 1, gray arrows show further references between articles, which are not considered by the network approach. The actual network structure is the structure of the German Wikipedia, article titles are translated into English.

3.3. Processing

Extracted coordinates from Wikipedia pages are provided by different Wikipedians. We have used the coordinate dump file of the German Wikipedia provided by the Wikipedia user 'Dispenser'.[3] The coordinates of these files are extracted on a daily basis. The dump file contains all coordinates that have been inserted into Wikipedia pages with the help of coordinate templates. However, the file does not only contain the headline coordinates, but also all other occurrences of coordinates within the full text of a Wikipedia page, for example, coordinates for the purpose of picture geo-tagging. As we decided to focus exclusively on coordinates in the headline, we post-processed the original dump file and tried to omit all articles without headline coordinates.

From our point of view, this step is essential for the network approach to geospatial reference, as inserting coordinates in the headline of a Wikipedia page is a conscious decision by the authors which indicates that these coordinates refer to the page as a whole, thus making the whole concept (geo)spatial. As the insertion of headline coordinates as well as of links into Wikipedia pages is of course a result of a cognitive process, the data analyzed by the network approach cannot be regarded as entirely 'non-cognitive'.

Besides checking for headline coordinates, we have excluded articles, which contained coordinates not referring to the earth, such as various pages on lunar craters and maria, because we wanted to concentrate on *geo*spatial references.

3.4. Results

The complete WAG, which is a result of the WikAPIdia tool, allows computing the NDGR for each article. Table 1 shows the results of the NDGR calculation for the German Wikipedia. The portion of geospatial articles is directly reflected by NDGR 0. It can be seen that only NDGR 0 to NDGR 3 are relevant. The 68 NDGR 4 articles are negligible. Articles that have NDGR higher than 4 do not exist. The cumulative proportions show that NDGR 0 and NDGR 1 amount to 78% of all articles. NDGR 0, NDGR 1, and NDGR 2 make up as much as 98.4%, which means that this portion of articles is not more than two clicks away from a page that has geo-coordinates.

Table 2 presents some typical examples of terms with NDGR 0, NDGR 1, and NDGR 2. We have not included NDGR 3 terms in this table as there were only two NDGR 3 terms in the whole survey. For this reason, we have also omitted NDGR 3 terms in some of the following figures. Note that geographic names are dominant but not exclusive for NDGR 0. Many of the NDGR 1 entities are either persons or terms that are strongly

Table 1. Results of NDGR calculations for the German Wikipedia. The Wikipedia XML dump file that was used is dated 21 June 2011, the coordinates file is dated 8 September 2011.

NDGR	Number of articles	Proportion (%)	Proportion, cumulative (%)
0	222,267	17.5	17.5
1	769,534	60.4	77.9
2	261,364	20.5	98.4
3	10,668	0.8	99.2
4	68	0.0	99.2
-1[a]	9,734	0.8	100.0
$\sum$	1,273,635	100.0	

Note: [a]NDGR $= -1$ indicates that the respective articles are not connected to any geospatial article. In almost all cases, this is because these articles are completely isolated.

Table 2. Some examples of terms of different NDGR (English translation via Interwikilinks).

NDGR 0	NDGR 1	NDGR 2
Cardiff	Neil Young	Career
Venezuela	Helmut Newton	Petroleum
University of Bremen	Battle of Cannae	Fantasy
CeBIT	Swiss People's Party	Public law
Havana Club	NHL Entry Draft	Cubic function

associated with geographic entities. Most of the NDGR 2 terms are rather abstract concepts. 'Neil Young',[4] for example, is an NDGR 1 entity because the corresponding article links to several places that were related to his life, for example, his birth place Toronto. 'Career',[5] on the other hand, is an NDGR 2 entity because among others the article links to the economist Richard Florida,[6] who himself is linked with Columbia University.

4. Cognitive approach to geospatial reference

4.1. Related work

While the network approach seems to be appropriate to analytically capture the 'geospatiality' of items within a network, it fails to tell us for which portion of information human cognition considers the geospatial reference as being relevant. For this reason, we also wanted to apply a cognitive approach and designed an experiment to shed more light on the question of how humans assess geospatial reference. By employing a cognitive approach, we align our research to the previous work that also applied cognitive approaches to geospatial phenomena, for example, Klippel *et al*. (2008, 2010), Freksa (1992), and Mark and Egenhofer (1994).

4.2. Categories of geospatial reference

As human cognition relies on categories, they need to be in the focus of the cognitive approach. Therefore, we have employed categories of geospatial reference (CGR) in our experiment: direct geospatial reference (DirGR), indirect geospatial reference (IndGR), and non-geospatial reference (NonGR). These three categories, which are elaborated in the literature (Bill 2010), may be artificial for non-experts. However, as their semantic structure is transparent, everybody is potentially able to interpret them in a non-arbitrary way. Thus, despite the artificiality of these categories, their interpretation is a cognitive process, which may also be seen as rating items on a three-point scale (Likert scale) for 'geospatiality'.

4.3. Experiment setting

Beside the main question, there are two further important issues that we wanted to address by the setting of our experiment: do results depend on (1) how we ask our participants and (2) on who we ask? The first question arises because biases introduced by the question method may be assumed, and the second question may be posed because a background of the participants in the field of geo-studies, geo-science, or geo-business may be expected to affect the results. In order to address these questions, we created two survey groups 'A' and 'B' that entailed two different formats of the survey questions and also recorded the geo-background of each participant.

4.3.1. Materials

In total, 1100 article titles have been randomly selected from the German Wikipedia considering two constraints. (1) The titles should not be ambiguous, such as 'Bank',[7] as this might result in an increased answer variance. As the respective pages contain a disambiguation hint in their headlines, this constraint was implemented with the help of the 'is_disambiguation' field, which is an output of the WikAPIdia tool. (2) The articles should rank among the top 5% of the most frequently accessed Wikipedia pages in 2011. This is because we assume that this ensures a maximum level of popularity as well as a comparably low portion of unknown terms. Access rates have been determined by analyzing the page count files provided by the Wikimedia foundation.[8] Any biases potentially introduced into the survey results by this method are not considered in this article, but could be taken into account in further research.

We have analyzed the survey results with respect to whether constraint (2) has introduced a bias. No significant correlations between the results of the experiment and the access rates were detected.

All the 1100 selected articles were used to create 11 distinct sets of 100 randomly selected terms. Based on these sets, 22 online surveys were generated, each set resulting in a pair of surveys referred to as set 'A' and set 'B' (cf. also Figure 3a and b).

4.3.2. Procedure

Participants were able to partake in the experiment via an online survey. They were free to choose when and where to do it. They received a short introduction explaining the context of the study. In particular, some citations of the 80%-assertion were referred to and it was mentioned that there is a lack of empirical justification in this field. The participants were not given any hint that, in a parallel study (network approach), a special focus was put on geo-coordinates in the headline of the Wikipedia articles.

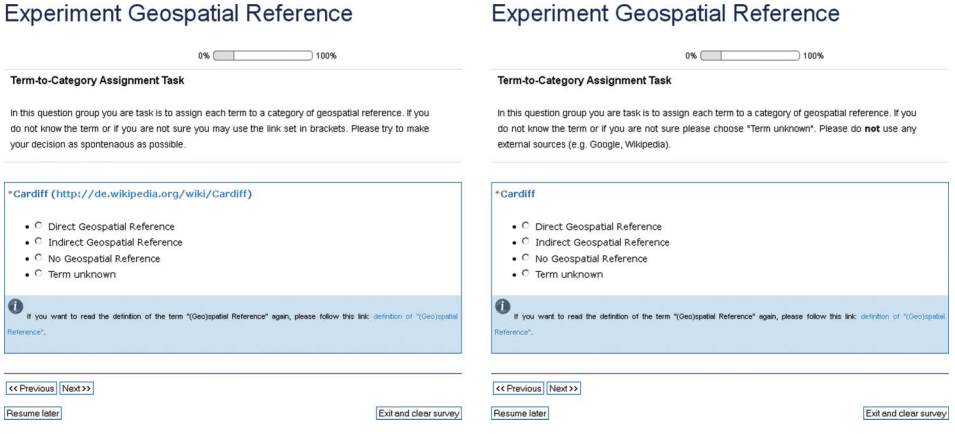

Figure 3. Example of the 'term-to-category assignment task' (English translation). (left) Set 'A' survey and (right) set 'B' survey. Group A includes the link to the corresponding article and also a note encouraging participants to use it, if they are unsure. Both the groups display a permanent link pointing to the given definition of geospatial reference. There are four possible categories for each term: DirGR, IndGR, NonGR, and 'Term unknown' (U). Note: Two surveys have been kept online for demonstration: group 1A: http://kartographie.geo.tu-dresden.de/limesurvey/64851/lang-de; group 1B: http://kartographie.geo.tu-dresden.de/limesurvey/68644/lang-de

Furthermore, participants were asked to enter anonymous personal data: (1) background in the field of geo-studies, geo-science, or geo-business yes/no; (2) awareness of the 80%-geo-assertion before participation yes/no; (3) age; (4) gender, and (5) current state of being a student yes/no. The task of the participants in the main experiment was to assign the titles of Wikipedia articles to the three CGR. In case they did not know a term or were not sure about it, they were asked to select 'term unknown'. For support, they were shown a slightly adapted version of the definition of geospatial reference by Bollmann (2002)[9] prior to the main experiment. It was made clear that there are no right or wrong ways for categorizing and that it was up to the participants to select criteria for categorizing themselves. No examples for categorization were given.

The terms in the main part of the experiment were presented in random order to reduce the influence of the participants' exhaustion. Figure 3 shows an example of the 'term-to-category assignment task'. Participants with a set 'A' survey were shown each term in combination with the link to the corresponding Wikipedia page. They were encouraged to follow it, if they felt they needed it to make their decision. Participants with a set 'B' survey were shown the same terms; however, they were not given any links and were asked not to use any external sources such as Google or Wikipedia.

4.3.3. Participants

In order to recruit participants, we posted an announcement on a German geo-news website.[10] Furthermore, we sent e-mail invitations to students, colleagues as well as other persons, whom the authors of the article know personally. Survey group selection for each participant was done by a randomized mechanism: all participants were given the same starting link, which redirected them.

4.4. Hypotheses

Our aim is to combine the results of the network and the cognitive approach in order to investigate the relationship between them. Our hypotheses are that (1) articles with a NDGR $= 0$ are predominantly categorized as DirGR, (2) articles with NDGR > 0 are rarely categorized as DirGR, but more frequently as IndGR, and (3) with increasing NDGR the portion of IndGR decisions is reduced in favor of an increasing portion of NonGR decisions. If our results confirm these hypotheses, then the two approaches support each other. If that is the case, the combination of both approaches is suitable to provide evidence for the initial geo-assertion though the 80%-figure might have to be adapted.

4.5. Data of experiment participation

In this Section, the background of the experiment participants is presented. Table 3 shows how many of the experiment participants had a background in geo-studies, geo-science, or geo-business and were aware of the 80%-figure. In the context of this study, it is remarkable that about 85% of all participants with a geo-background have heard about the 80%-figure. However, this might not be fully representative, since people who knew this figure were more likely to participate. In general, the study met with a considerable eagerness among geo-folk to participate. Consequently, the share of geo-background participants is rather high. With regard to those participants, who knew the figure without having a geo-background themselves (2.6%), we have to concede that their knowledge is a direct consequence of their acquaintance with the authors of this article.

Table 3. Portions of participants with a geo-background (studies, business, or scientific): Geo = yes/Geo = no and portion of participants who have heard about the 80%-figure before participation.

	Geo = yes (total)	Geo = yes (%)	Geo = no (total)	Geo = no (%)	∑ (total)	∑ (%)
'80%' = yes	245	64.5	10	2.6	255	67.1
'80%' = no	47	12.6	78	20.5	125	32.9
∑	292	76.9	88	23.1	380	100.0

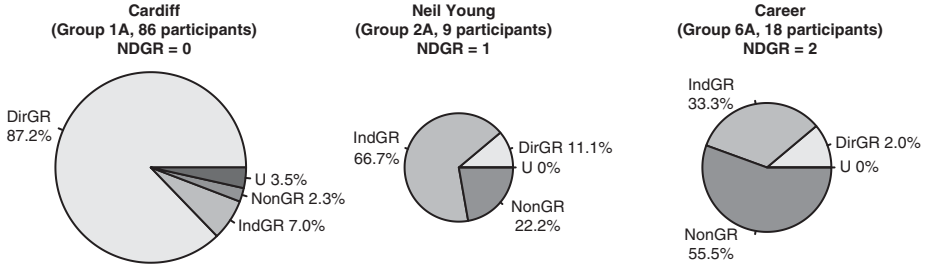

Figure 4. Pie charts showing different portions of CGR decisions for the articles 'Cardiff', 'Neil Young', and 'Career'. The numbers of participants are indicated by different sizes of the pies.

In total, 380 people participated in the experiment. At the time of participation, 33% of them were students. The average age was 34.5 years. The portion of female/male participants was 37%/63%. For reference, we have kept the results of all survey groups online.[11]

4.6. Results

For the cognitive approach, we computed for each article the portions of the different CGR that had been assigned by the participants. Figure 4 shows three example articles with different NDGR. For 'Cardiff', 75 participants selected 'DirGR', six participants selected 'IndGR', two participants selected 'NonGR', and three participants selected 'U' (term unknown). This results in 87.2% 'DirGR', 7.0% 'IndGR', 2.3% 'NonGR', and 3.5% 'U'. For the analyses, it will be assumed that participants that selected 'U' would have selected any category with the same probability as all other participants, if they had known the term. Consequently, portions of CGR decisions may be computed relative to the portion of all not-'U' decisions. For the case of the Cardiff example, this results in these adapted portions of CGR: 90.4% 'DirGR', 7.2% 'IndGR', and 2.4% 'NonGR'. Complete results for all articles that were tested in the survey can also be found in the online survey statistics.

5. Synthesis of approaches

5.1. Methodology

In this section, the interplay between the network approach and the cognitive approach will be discussed. Our main focus is to answer the question to which categories participants assigned terms with different NDGR. For this purpose, we synthesize the results of the network approach and the cognitive approach in the following way: we combine the NDGR

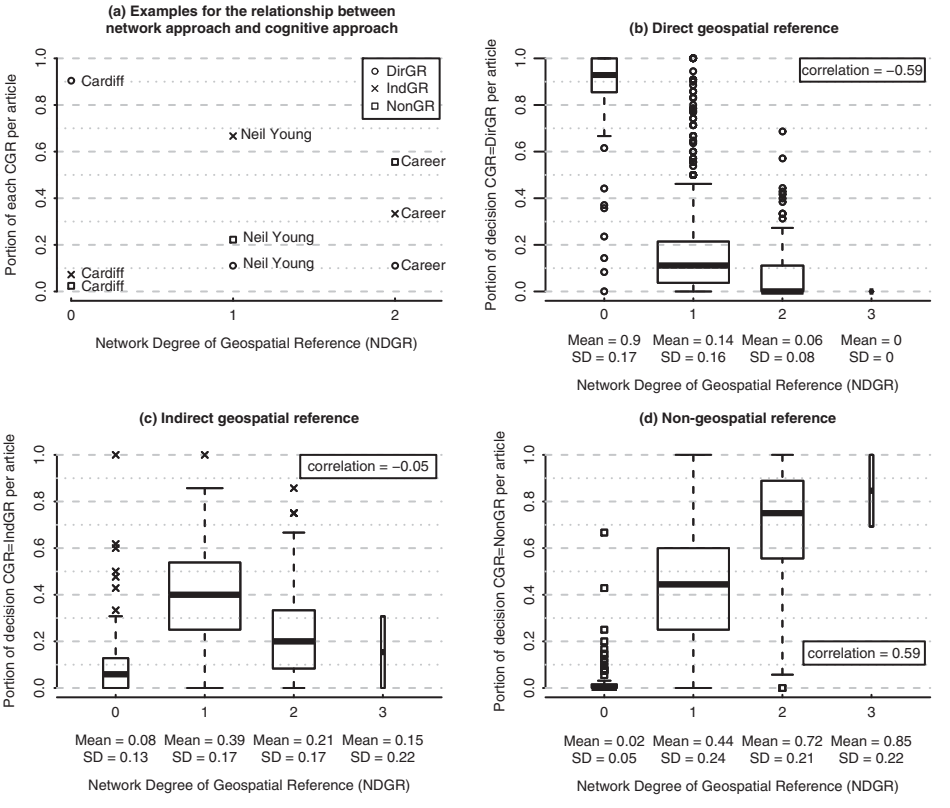

Figure 5. (a) The portions of DirGR, IndGR, and NonGR as well as the NDGR of the three example articles (cf. Figure 3) transferred to an *x*-*y* plot. (b)–(d) Box-and-whisker plots for the relationship between NDGR and different CGR. Results of group A (participants encouraged to use Wikipedia links). In total, 1100 articles have been investigated.

of a specific article and the portion of participants that assigned a specific CGR to the title of this article into one two-dimensional value. As there are three CGR portions per article, three values can be generated per article. Figure 5a shows a plot of the nine values generated from the three-example articles shown in Figure 4, among others the values of 0/0.87 (NDGR and portion of DirGR for 'Cardiff'), 1/0.67 (NDGR and portion of IndGR for 'Neil Young'), and 2/0.056 (NDGR and portion of NonGR for 'Career'). Figure 5b–d visualizes the relationship between the network approach and the cognitive approach for the values generated from *all* articles in the form of box-and-whisker plots (cf. Tukey 1977).

In this article, we scale the width of the boxes according to the number of articles represented by each box. In order to reduce the impact of outliers, we computed robust mean values as symmetrically trimmed means with a fraction of 0.05 observations deleted from each end of the distribution (cf. Wilcox 2004, pp. 56–59). Likewise, we used a robust estimator for the calculation of standard deviations. Wilcox (2004, pp. 62–63) discusses the sample Winsorized variance, which we applied with 0.05 as the amount of Winsorization. The square root of the sample Winsorized variance was used to estimate standard deviations. For the computation of correlation coefficients, we applied Pearson's correlation coefficient.

Figure 5b shows the relationship between NDGR and DirGR. As expected, there is a negative correlation (–0.59) between the portion of DirGR per article and the NDGR of an article. NDGR 0 articles have been categorized by the participants as DirGR with an average of 90%. This high matching rate shows that the results support each other, which provides evidence for two things: for human cognition 'direct geospatial reference' is available as a category and, secondly, the network approach is a suitable method to measure geospatial reference. Upper outliers of NDGR 1 might be candidates for articles that are still lacking a headline coordinate.

Figure 5c shows the relationship between NDGR and IndGR. As can be seen, there is no linear correlation. Instead, there is a non-linear relationship with a peak at NDGR 1. The portion of NDGR 0 is low because NDGR 0 articles have mainly been assigned to DirGR. NDGR greater than 1 show a constant decrease of IndGR portions.

Figure 5d shows the relationship of NDGR and NonGR which is similar to the inverse relationship between NDGR and DirGR. The higher the NDGR, the higher is the portion of NonGR, which results in a positive correlation between NDGR and NonGR of 0.59. The height of the boxes and the length of the whiskers indicate the dispersion of answers. It can be seen that the height of the NDGR 0 boxes is small for all three categories. From this, it may be inferred that categorization of NDGR 0 terms is relatively unambiguous. As the boxes of NDGR 1 and NDGR 2 for the categories IndGR and NonGR are particularly high, we may conclude that the category of IndGR is rather fuzzy and that the threshold that delineates IndGR and NonGR is fuzzy, too. In summary, we found that all three hypotheses that were formulated in Section 4.4 to be confirmed.

5.2. Impact of experimental method on the categorization results

As mentioned earlier, we wanted to investigate the influence of 'how' we asked the participants and of 'who' we asked. Figure 6a and b illustrate how different factors influenced the results.

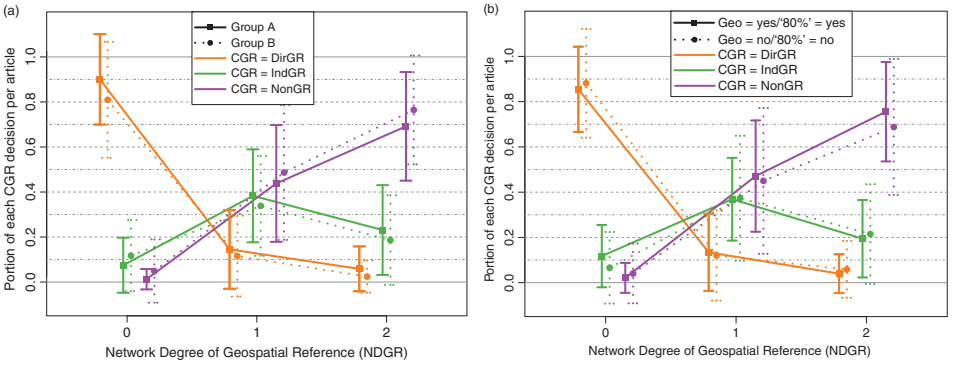

Figure 6. Error bar plots. Lengths of error bars represent standard deviations. (a) Comparison of results of group A and group B. Group A participants were encouraged to use Wikipedia article contents. Group B participants were discouraged to use any external sources. As there were less group B participants (152 in total), for each set of 100 questions a random sample of group A participants was selected, resulting in a similar number of group A participants (157). (b) Comparison of geo-participants and non-geo-participants. A participant is considered a geo-participant if he/she indicated to have a geo-background and to have known the 80%-figure before participation.

Figure 6a shows the results of group A participants in comparison to group B participants. Participants in group A were encouraged to gather further information before making their decisions. Participants in group B were asked not to use any external sources.

As can be seen in Figure 6a, there is a direct impact of the applied method on the results of the categorization task. Participants that were allowed to inform themselves and hence had some sort of activated pre-knowledge tended to categorize terms more spatially. The mean values connected by the orange lines show this for DirGR. Group A results (solid lines) occur above group B results (dotted lines) for all NDGR. Likewise, the IndGR mean values (green) of group A are above the corresponding mean values of group B, except for NDGR 0. The results for NonGR (purple) correspond to the other results as they show that the portion of participants that decided to choose the 'non-geospatially referenced' category is lower for group A than for group B. Nevertheless, even the fact that some of the differences between both groups are statistically significant is not enough evidence to conclude that activated pre-knowledge changes the results completely. Beside the differences in the mean values, we may observe that standard deviations in group A are generally lower than in group B. From that we can deduce that participants who were encouraged to take further information into account produced more stable results.

5.3. Impact of professional background on the categorization results

The results of the participants are compared with regard to their professional background in Figure 6b. In order to be able to apply the best possible distinction between both groups of participants, we decided to consider only those with a geo-background and a knowledge of the 80%-figure, on the one hand, and those that had neither, on the other hand.

In general, we may notice that the differences between geo-participants and non-geo-participants are smaller than the differences between group A and group B participants. From this, we may infer that it is less important 'who' we ask than 'how' we ask. Furthermore, these results provide evidence that the geo-participants do not generally consider the tested Wikipedia articles to be more geospatially referenced than non-geo-participants. Hence, it seems that people with a geo-background do not have a particularly 'geospatially influenced' worldview. However, some differences between both groups of participants may be observed: for NDGR 0 articles the non-geo-participants have less frequently chosen IndDR and more frequently either DirGR or NonGR. This may indicate that, for the non-geo-participants, the category of IndGR is less clear than DirGR and NonGR. However, this is not confirmed by the results of NDGR 1 and NDGR 2. Apart from that, the results of NDGR 2 show that the geo-participants have more frequently chosen NonGR. The reason for that may be that the geo-participants have a stronger idea of what is geospatially referenced and what is not.

5.4. Prediction of CGR portions for the whole German Wikipedia

The research work that we present in this article was triggered by the question how much information is geospatially referenced. Consequently, we computed a prediction of the portions of DirGR, IndGR, and NonGR for the whole German Wikipedia based on the results of the network approach and the cognitive approach. Table 4 shows the predicted values. The mean values of each NDGR-CGR combination are the same as in Figure 4b–d. We have estimated the standard errors of the mean values according to Wilcox (2004, p. 63) applying 2σ accuracy. The portions of NDGR for the whole German Wikipedia are adapted values of Table 1 – we have computed these values relative to all non-isolated articles (cf. Table 1: NDGR $= -1$). As NDGR 4 is negligible, we have omitted

Table 4. Prediction of CGR portions for the whole German Wikipedia.

CGR	NDGR	0 17.6%	1 60.9%	2 20.7%	3 0.8%	∑
DirGR		0.90 ± 0.04 *15.2% ± 0.7%*	0.14 ± 0.01 *10.4% ± 0.8%*	0.06 ± 0.01 *1.4% ± 0.2%*	0.00 ± 0 *0.0% ± 0%*	**27.0% ± 1.1%**
IndGR		0.08 ± 0.03 *2.0% ± 0.5%*	0.39 ± 0.01 *23.9% ± 0.9%*	0.21 ± 0.02 *4.4% ± 0.5%*	0.15 ± 0.34 *0.1% ± 0.3%*	**30.3% ± 1.2%**
NonGR		0.02 ± 0.01 *0.7% ± 0.2%*	0.44 ± 0.02 *26.9% ± 1.2%*	0.72 ± 0.03 *14.3% ± 0.6%*	0.85 ± 0.34 *0.8% ± 0.3%*	**42.7% ± 1.4%**

it. For each NDGR-CGR combination, the percentage of its NDGR portion and the mean value/standard error of its CGR portion were multiplied (results in italics). The total portion of each CGR (in bold) is the sum of all mean values of this CGR. The total standard error of each CGR (in bold) is computed using the Bienaymé formula (cf., e.g., Loève 1977, p. 12). As discussed in Section 4.5, the results of group A have lower standard deviations. Since this indicates more stable decisions of the participants, we have decided to use these results for the prediction.

The predicted portion for the category of DirGR is 27.0% (±1.1%). For the category of IndGR, it is 30.3% (±1.2%). Together both CGR make up a portion of 57.3% (±1.6%) compared to a portion of 42.7% (±1.4%) for non-geospatially referenced information. Despite the comparably low standard errors, ranges rather than point estimations may be preferred as results. Consequently, we take the 2σ errors as interval estimators to deduce the following ranges as results: 26–28% DirGR, 29–32% IndGR (DirGR + IndGR: 56–59%), and 41–44% NonGR.

6. Discussion

The drawback of the presented approaches in conjunction with the employed corpus of a specific Wikipedia (i.e., an encyclopedia) is that they fail to estimate portions of geospatial reference when focusing on concrete instances instead of generic concepts. This is because it is neither possible to infer the categorization of single instances from the categorization of concepts nor to estimate the number of instances for most concepts. While, for example, the generic concept ‘house’ might not be considered to have a direct geospatial reference *per se*, a concrete instance of a house, such as the ‘Empire State Building’ or the ‘White House’, may of course be considered to be directly geospatially referenced. Furthermore, it is not always possible to estimate how many instances of a concept there are in total. Moreover, there are concepts that do not have any physical instances at all, such as ‘philosophy’ or ‘synthpop’. In order to focus on instances, a preferably domain-neutral, graph-structured corpus of single data entities is required. For this purpose, the Semantic Web may be suitable.

Furthermore, the network approach presented in this article neither accounts for the specific number of links that connect NDGR 1, NDGR 2, or NDGR 3 concepts

to geospatial concepts nor for different types of links, which may be distinguished in other corpora than the employed German Wikipedia. However, both factors may be of importance for human cognition. Consequently, refinements to the network approach should yield results that are closer to human categorization of geospatial reference.

One suggestion for such a refinement would be the use of weights for NDGR calculation. These weights can be based on link counts and also on link semantics. In both cases, we would need to switch the employed graph model from a directed *unweighted* graph to a directed *weighted* graph.

Link counts may easily be integrated as they may be gained by using the WikAPIdia tool. With regard to the semantics of links, weights would need to be chosen with respect to different types of links depending on how they constitute geospatial reference. In this regard, it needs to be considered that there are different kinds of transformations that make something geospatially referenced, such as address-to-coordinate conversion (geocoding), the use gazetters, or sensors (e.g. GPS devices). The Semantic Web may be a suitable way to integrate the semantics of these references, because it explicitly models them. In the ontology of DBpedia, there are properties, such as dbo:location, dbo:birthplace, dbo:hometown, and dbo:premierePlace.[12] However, further investigations would be required to define criteria that might be used to derive weights for different types of links. The large number of link types would make this an even bigger challenge.

The cognitive approach may also be refined. One such refinement could be realized by a 'group-and-rank' task. Participants could be asked to create groups for terms whose 'geospatiality' is perceived as similar. Secondly, they would have to rank these groups. This would help us to find terms that human cognition considers more geospatially referenced than others. Moreover, it would be possible to analyze for each article how long it took participants to assign it to a CGR. These data have already been collected in our study. This aspect might provide additional evidence, especially for the estimation of the uncertainty of decisions.

As discussed earlier in this article, we do not expect significant differences depending on the Wikipedia language version for the network approach. However, there may be linguistic effects on the conceptualization of the CGR that may have an impact on the results of the cognitive approach. In this context, Klippel and Montello (2007) show how language affects the human conceptualization of directions. Consequently, there would be a demand for a subsequent study employing another language version of Wikipedia.

7. Conclusions and future work

The work we have presented in this article is a contribution to the quantification of geospatially referenced information within user-generated content. It may serve as a proxy for the real situation, which would be much harder to investigate. Moreover, the network approach may be used to estimate the 'geospatiality' of information in the context of semantic or geospatial web analyses. In summary, it can be said that the synthesis of the two presented approaches – network approach and cognitive approach – is suitable to estimate the portion of geospatially referenced items within graph-structured corpora of information.

According to the results of our study, 57% of the information within the German Wikipedia is geospatially referenced. This total amount of geospatially referenced information consists of a 27% share that is categorized by humans as information with direct geospatial reference and a 30% share that is categorized as information with indirect geospatial reference.

Although this is below the original estimations of the geo-community, which may have been biased by their wishes, it is still strong evidence for the relevance of the geospatial

information community. Hence, from the authors' point of view, the results of this study do not need to be seen as detrimental to the standing of geoinformation business and geoscience. It is certainly preferable to refer to a lower number that is supported by a scientific investigation than to postulate a higher number that does not seem to be reliable. However, we also have to bear in mind that these results should not be overgeneralized. The original assertion referred to municipal and government data and our study might have produced a higher percentage, if we had investigated this type of data only. In contrast to that, our results have been gained using Wikipedia (i.e. an encyclopedia), which might be considered as domain-neutral knowledge. However, it needs to be considered that Wikipedia is a product of a rather small number of authors (Ortega *et al*. 2008). Hence, our method strongly depends on the work of this minority of Wikipedians. In summary, we have shown that the very general context-free assertion that 80% of all information is geospatially referenced may be falsified.

Acknowledgements

Research for this article is based upon work supported by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) under the project 'Mobile map applications based on user generated content for cartographic communication' (BU 2605/1-1). We acknowledge the support of Nicolas Chrisman in tracing the references of the origins of the assertion. We furthermore thank Brent Hecht, who implemented the WikAPIdia library that we have used to parse the German Wikipedia. The work of various Wikipedia community members, who have put significant effort into producing the coordinate files, is highly appreciated. We are grateful to 380 people who participated in our survey. We sincerely thank Beatrix Weber and five anonymous reviewers for their valuable comments to the manuscript and their constructive suggestions.

Notes

1. x, y, z, and t denote coordinates within space–time; U denotes an arbitrary thing or property.
2. Within this simulation, results were calculated after omitting the headline coordinates from a random 33% sample of the geospatial articles, which resulted in a lower coverage of geospatial articles (11.6%), see also Sections 3.2 and 3.4. Because of the dense network structure of the examined German Wikipedia, the results of this simulation were similar to the results shown in Table 1.
3. http://toolserver.org/~dispenser/dumps/
4. Cf. http://en.wikipedia.org/wiki/Neil_Young
5. Cf. http://en.wikipedia.org/wiki/Career
6. Cf. http://en.wikipedia.org/wiki/Richard_Florida
7. Cf., for example, disambiguation of "Bank": http://en.wikipedia.org/wiki/Bank_(disambiguation)
8. Page count files are generated from server access logs, for the study files from the period between September 2010 and August 2011 were analyzed. Files can be downloaded from: http://dumps.wikimedia.org/other/pagecounts-raw/
9. http://kartographie.geo.tu-dresden.de/geospatial_experiment/definition.htm
10. http://www.geobranchen.de/index.php?option=content&task=view&id=5053
11. http://kartographie.geo.tu-dresden.de/geospatial_experiment/results.htm
12. Namespace of the DBpedia ontology: http://dbpedia.org/ontology/ (dbo)

References

Bill, R., 2010. *Grundlagen der Geoinformationssysteme*. Berlin: Wichmann.

Bizer, C., *et al*., 2009. DBpedia – a crystallization point for the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7 (3), 154–165.

Bollmann, J., 2002. Raumbezug. *In*: J. Bollmann and W.G. Koch, eds. *Lexikon der Kartographie und Geomatik*. Heidelberg: Spektrum Akademischer Verlag, 266.

Broder, A., *et al*., 2000. Graph structure in the Web. *Computer Networks*, 33 (1–6), 309–320.

Capocci, A., *et al*., 2006. Preferential attachment in the growth of social networks: the internet encyclopedia Wikipedia. *Physical Review E*, 74 (3), 36116.

Cardoso, N., 2011. Evaluating geographic information retrieval. *SIGSPATIAL Special*, 3 (2), 46–53.

Dahinden, T., 2011. Estimation of the locations of the language-versions of Wikipedia – a case study on geographic data mining. *In*: A. Ruas, ed. *Advances in cartography and GIScience: selection from ICC 2011, Paris*. Berlin: Springer, 471–487.

Fitzke, J. and Greve, K., 2010. Frei oder umsonst? – Nutzergenerierte Geoinformation zwischen Freiheit und Kostenlosigkeit. *In*: J. Strobl, T. Blaschke, and G. Griesebner, eds. *Angewandte Geoinformatik 2010: Beiträge zum 22. AGIT-Symposium*. Berlin: Wichmann, 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), 10–22.

Freksa, C., 1992. Using orientation information for qualitative spatial reasoning. *In*: A. Frank, I. Campari, and U. Formentini, eds. *Theories and methods of spatio-temporal reasoning in geographic space*. Berlin: Springer, 162–178.

Goodchild, M.F., 2003. The nature and value of geographic information. *In*: M. Duckham, M.F. Goodchild, and M. Worboys, eds. *Foundations of geographic information science*. New York: Taylor & Francis, 18–30.

Goodchild, M.F., *et al*., 1999. Introduction to the Varenius project. *International Journal of Geographical Information Science*, 13 (8), 731–745.

Haggett, P., 2001. *Geography: a global synthesis*. Harlow: Prentice Hall.

Hahmann, S., Burghardt, D., and Weber, B., 2011. "80% of all information is geospatially referenced"??? Towards a research framework: using the semantic Web for (in)validating this famous geo assertion. *In*: *AGILE Paper Sessions*, Universiteit Utrecht.

Hecht, B. and Gergle, D., 2010. The Tower of Babel meets Web 2.0: user-generated content and its applications in a multilingual context. *In*: E. Mynatt, *et al*., eds. *CHI '10: proceedings of the 28th international conference on human factors in computing systems*. New York: ACM, 291–300.

Hecht, B. and Moxley, E., 2009. Terabytes of Tobler: evaluating the first law in a massive, domain-neutral representation of world knowledge. *In*: K.S. Hornsby, *et al*., eds. *Spatial information theory: 9th international conference. LNCS 5756*. Berlin: Springer, 88–105.

Huxhold, W.E., 1991. *An introduction to urban geographic information systems*. New York: Oxford University Press.

Klippel, A. and Montello, D.R., 2007. Linguistic and nonlinguistic turn direction concepts. *In*: S. Winter, M. Duckham, L. Kulik and B. Kuipers (eds.), *Proceedings of the 8th international conference on spatial information theory (COSIT)*, 19–23 September, Melbourne, Australia. Berlin: Springer-Verlag, 354–372.

Klippel, A., Worboys, M., and Duckham, M., 2008. Identifying factors of geographic event conceptualisation. *International Journal of Geographical Information Science*, 22 (2), 183–204.

Klippel, A., *et al*., 2010. Cognitive invariants of geographic event conceptualization: what matters and what refines? *In*: S. Fabrikant, *et al*., eds. *Geographic information science*. Berlin: Springer, 130–144.

Lawrence, V., 2009. *The role of a national mapping agency in geoinformation management* [online]. Southampton, Ordnance Survey. Available from: http://www.fig.net/pub/fig2009/ppt/ps01/ps01_lawrence_ppt_3505.pdf [Accessed 20 Mar 2012].

Loève, M., 1977. *Probability theory I*. 4th ed. New York: Springer.

Mark, D.M., 2003. Geographic information science: defining the field. *In*: M. Duckham, M.F. Goodchild, and M. Worboys, eds. *Foundations of geographic information science*. New York: Taylor & Francis, 3–18.

Mark, D.M. and Egenhofer, M.J., 1994. Modeling spatial relations between lines and regions: combining formal mathematical models and human subjects testing. *Cartography and Geographical Information Systems*, 21 (3), 195–212.

McCurley, K.S., 2001. Geospatial mapping and navigation of the web. *In*: V. Shen, N. Saito, M.R. Lyu and M.E. Zurko (Eds.), *Proceedings of the 10th international conference on World Wide Web*, 1–5 May, Hong Kong. New York: ACM, 221–229.

Municipality of Burnaby, 1986. *Invitation to information*. Burnaby, Canada: Brochure.

Ortega, F., Gonzalez-Barahona, J., and Robles, G., 2008. On the inequality of contributions to Wikipedia. *In*: R.H. Sprague Jr. (Ed.), IEEE Computer Society, ed. *Proceedings of the 41st*

annual Hawaii international conference on system sciences, 7–10 January, Waikoloa, HA. Los Alamitos, CA: IEEE Computer Society, 304–310.

Perkins, B., 2010. *Hav you mapped your data today?* [online]. Framingham, Computerworld. Available from: http://www.computerworld.com/s/article/350588/Have_You_Mapped_Your_Data_Today_ [Accessed 10 Oct 2012].

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*: M. Konecny, ed., *Proceedings of the 4th AGILE conference on GIScience*, 19–21 April 2001 Brno, Czech Republic. AGILE, 215–221.

Shannon, C.E. and Weaver, W., 1949. *The mathematical theory of communication*. Urbana, IL: University of Illinois Press.

Tukey, J.W., 1977. *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Voss, J., 2005. Measuring Wikipedia. *In*: P. Ingwersen and B. Larsen, eds. *Proceedings of the 10th international conference on scientometrics and informetrics*, 24–28 July 2005, Stockholm, Sweden. Available online at: http://eprints.rclis.org/handle/10760/6207

Watts, D.J. and Strogatz, H., 1998. Collective dynamics of 'small-world' networks. *Nature*, 393 (6684), 440–442.

Wilcox, R.R., 2004. *Introduction to robust estimation and hypothesis testing*. 2nd ed. San Diego, CA: Academic Press.

Zlatić, V., *et al*., 2006. Wikipedias: collaborative web-based encyclopedias as complex networks. *Physical Review E*, 74 (1), 16115.