Extracting central places from the link structure in Wikipedia

Kessler, Carsten

Published in:
Transactions in G I S

DOI (link to publication from Publisher):
10.1111/tgis.12284

Publication date:
2017

Document Version
Early version, also known as pre-print

Citation for published version (APA):
Kessler, C. (2017). Extracting central places from the link structure in Wikipedia. *Transactions in G I S*, *21*(3), 488-502. https://doi.org/10.1111/tgis.12284

This is a preprint version of:
Carsten Keßler (2017) Extracting Central Places from the Link Structure in Wikipedia. Transactions in GIS 21(3):488–502. DOI:10.1111/tgis.12284

Extracting Central Places from the Link Structure in Wikipedia

Carsten Keßler
kessler@plan.aau.dk
Department of Planning
Aalborg University Copenhagen

Abstract

Explicit information about places is captured in an increasing number of geospatial datasets. This paper presents evidence that relationships between places can also be captured implicitly. It demonstrates that the hierarchy of central places in Germany is reflected in the link structure of the German language edition of Wikipedia. The official upper and middle centers declared based on the German spatial laws are used as a reference dataset. The characteristics of the link structure around their Wikipedia pages – which pages link to each other or mention each other, and how often – are used to develop a bottom-up method to extract central places from Wikipedia. The method relies solely on the structure and number of links and mentions between the corresponding Wikipedia pages; no spatial information is used in the extraction process. The output of this method shows significant overlap with the official central place structure, especially for the upper centers. The results indicate that real-world relationships are in fact reflected in the link structure on the web in the case of Wikipedia.

1 Introduction

Places, structures and activities in the real world are captured as data in a number of different ways. They range from datasets produced by remote sensing techniques and professional mapping activities to location information in social networks and volunteered geographic information. These examples show that there are many datasets that reflect our real-world environment from different angles. The research presented in this paper adds another dimension to this insight by demonstrating that the relationships between places in the real world are reflected by the structure of the web. More specifically, it points out similarities between the place hierarchy in a country and the link structure between the Wikipedia pages about those places. The status of a city as a center for production, trade, and administration, for example, is reflected in a large number of incoming links to its Wikipedia page. Its importance for places within its range – or rather, the people living there – leads to a significantly higher number of incoming links from pages about those nearby places. These two characteristics allow us to infer the hierarchy of places in a country from the link structure in Wikipedia.

Christaller attempted to explain the arrangement of and relationships between places in his Central Place Theory (CPT) (Christaller, 1933; Baskin, 1966). It predicts that under perfect conditions, settlements of different orders arrange in a hexagonal pattern, with lower-order settlements such as towns forming a hexagon around one higher-order settlement such as a city, which provides goods, services, and jobs for the lower-order places in its range. In an earlier publication by this author, it has been shown that these hexagonal patterns can be derived from the link structure in the English language Wikipedia to a certain extent (Keßler, 2015). However, it faced the problem that place structures in reality hardly ever expose the hexagonal arrangement predicted by CPT because of topography and uneven distributions of resources and population, to name but a few reasons.

Therefore, this paper takes a more systematic and more realistic approach by comparing the link structure in the German language Wikipedia to the *actual* place hierarchy in Germany. Since the German spatial planning laws require the determination of three different levels of centers, this official place hierarchy provides a valuable reference point for this study. It eliminates the need to somehow assess the centrality of a place, which is often impossible without local knowledge and a consideration of *all* places in an area. The reference dataset generated from the German place hierarchy takes into account the first two levels and includes a total of 997 upper and middle centers. It enables a more systematic, quantitative analysis that goes beyond the qualitative assessment in Keßler (2015) and allows for more generic insights about the way real-world structures are reflected in what can be seen as a special kind of Volunteered Geographic Information (Goodchild, 2007).

The research question addressed in this paper is therefore whether and to what extent the link structure in Wikipedia reflects a real-world network of central places. It is answered by (1) analyzing the properties of an official network of central places, and the Wikipedia pages of the corresponding centers; (2) using those properties to define a bottom-up extraction approach to identify centers in Wikipedia; and (3) comparing the centers identified by this approach to the official central places. The results of the paper show that using just a single feature of Wikipedia – namely the number of links and mentions between pages about cities – is sufficient to reconstruct a real-world network of central places to a large degree.

The remainder of the paper is organized as follows: The next section presents relevant related work, followed by an introduction to the datasets used in this research in Section 3. Section 4 provides different perspectives on the characteristics of Wikipedia pages about central places in Germany. Section 5 uses these characteristics to extract a central place structure from Wikipedia and compares the result to the official central places, followed by concluding remarks in Section 6.

2 Related Work

This section provides an overview of relevant related work on Central Place Theory and the analysis of spatial aspects of the web and Wikipedia.

2.1 Central Place Theory

Christaller introduced Central Place Theory (CPT) as an explanation of the spatial arrangement of places as economic centers that people visit to work or trade (Christaller, 1933; Baskin, 1966). Assuming an isotropic plane as well as evenly distributed population and resources, it states that places will arrange in a hexagonal structure of five different levels of centers, from small hamlets up to large regional capitals, each with different functions offered to the people in their range. Later revisions of the theory moved away from a strict economic perspective and took consumer welfare into account (Lösch, 1954). The predicted hexagonal spatial configurations of places have been compared to the real spatial configurations of places by different researchers (Brush, 1953; Berry and Garrison, 1958b, e.g.), confirming that the theory does predict real-world situations well when the local conditions are close to the (often unrealistic) assumptions underlying CPT.

More recent research has compared the trip distributions between places that CPT predict to different spatial interaction models and finds significant differences (Openshaw and Veneris, 2003). Hsu (2012) has investigated city size distributions between different kinds of centers and found that they follow a power law distribution. Keßler (2015) presents a qualitative analysis of central place structures for large cities in the US derived from the English language Wikipedia. The paper discusses these structures' similarities to the structures predicted by CPT. While the results are plausible, the results remain on the level of a qualitative discussion because there is no official declaration of centers for the US, and the actual spatial configuration of those centers often deviates significantly for the predictions by CPT.

While the tenets of CPT are not always supported by observations in reality, the arrangement of places around centers can be observed everywhere in the world. Germany has therefore decided to use different levels of centers and their ranges as a planning instrument, and requires their declaration in the states' spatial structure plans (Raumordnungsgesetz, 2015). The upper two of the three levels of centers declared on the basis of this law will be used as a basis for the analysis in this paper.

2.2 Geographic Aspects of the Web and Wikipedia

Geographic information on the web comes in many different forms, including different flavors of volunteered geographic information. Previous research has focused on using this information in a variety of ways, including for question answering (Santos and Cardoso, 2008), similarity-based place search (Adams and McKenzie, 2012), the approximation of feature outlines (Keßler et al., 2009), and even the construction of whole gazetteers (Keßler et al., 2009; Gao et al., 2017), among others. Wikipedia has already been the focus of many research efforts, most likely due to free and easy access to the rich dataset provided by the community-driven encyclopedia. Overell and Rüger (2006) show that Wikipedia can be used to significantly improve the disambiguation of place names. Takahashi et al. (2011) present an approach to infer the significance of spatio-temporal events from Wikipedia; like in this research, links play a major role in their analysis of events. Hardy (2010) investigates the spatial behavior in the editing of Wikipedia articles. His results show that contributors edit articles

about nearby places, with an exponential decay in the influence of proximity. Moreover, Lieberman and Lin (2009) state that the geographic coordinates of pages edited by a user often cluster tightly. Hecht and Gergle (2010), however, point out that Wikipedia content is less 'local' than other forms of volunteered geographic information, such as geotagged photos.

Among the research dealing with geographic aspects of Wikipedia, two previous studies are particularly close in nature to the research presented here. Hecht and Moxley (2009) demonstrate the validity of Tobler's First Law of Geography (Tobler, 1970) in various language editions of Wikipedia. Their experiment shows that the Wikipedia pages of nearby places are more likely to link to each other than those of distant pages. The research conducted by Salvini (2012) analyzes the structure of the global network of cities based on the English language edition of Wikipedia. The research makes use of spatialization techniques to identify networks for different kinds of functions, such as politics, education, or art.

3 Datasets

This section introduces the dataset used in this study and the reference dataset containing an actual central place structure. Moreover, it explains the steps required to prepare the data for efficient analysis.

3.1 German Wikipedia Data

The dataset used in this study is based on the German Wikipedia as of February 2nd, 2016,[1] which consists of 2,382,442 pages. The dataset dump used here (see Section 3.3) contains the full text for all pages, including links to any other pages. Internal links that point to other pages on the German Wikipedia and the corresponding *mentions* are a central element for the analysis of central place structures presented here. Mentions are defined as any text element that also appears as a link on the same page, as shown in Figure 1: If at least one 'clickable' hypertext link from page A to page B is found, page A is scanned for any other occurrences of B in its text that are not hyperlinks. Since links in the text of a wikipedia page are explicitly set by a human editor, it is safe to assume that topic B has some relevance in the context of topic A if such a link has been set. The rationale behind the mentions is that Wikipedia editors of page A usually only include one hyperlink to page B, and leave all (or most) other mentions of B as plain text. In order to assess the degree of relevance of page B in the context of page A, such mentions need to be considered. In this research, the count of links and mentions is used as a proxy to estimate how important one place is for another one. Between the ∼2.3 million pages in the Wikipedia dump used here, there are 73,283,735 unique pairs of pages with at least one link from one to the other. Overall, these add up to 91,143,727 links and 2,114,568,525 mentions, for a total of ∼2.2 billion references.

The second element used in this research is the geotag from each page that contains one (see Figure 2). The geotags are available as a separate dataset which contains point coordinates for 364,150 pages in the German Wikipedia. This dataset also contains a coarse classification of the entities described on

[1] Obtained from `https://dumps.wikimedia.org/dewiki/`.

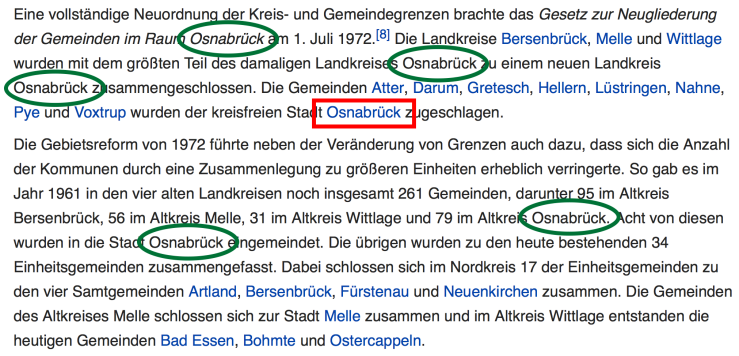

Figure 1: The red box shows the link to the page *Osnabrück*. The green ellipses highlight all mentions of the same page.

the corresponding pages, along with information about the country and state. Overall, 44,427 pages in the dump used here are classified as cities[2] in Germany. Out of the total ∼73 million pairs of pages with at least one link between them, there are 3,079,115 pairs where both pages are geotagged, which add up to 3,783,421 links and 11,998,844 mentions, for a total of ∼15.8 million references.

Figure 2: Example of a geotag in a Wikipedia page.

3.2 Reference Dataset

The declaration of central places at different levels is an important planning instrument in Germany. According to the German Raumordnungsgesetz (2015, §8), the individual states (*Länder* in German) have to develop spatial structure plans that can contain information about central places. In practice, they declare lower, middle, and upper centers (*Unterzentren*, *Mittelzentren*, and *Oberzentren* in German) as part of their respective state development plans. While the terminology and definitions of those different classes of centers are not entirely consistent between all states, the general idea is that lower centers

[2]Besides *city*, the classification system used here only contains the classes *adm1st*, *adm2nd*, *airport*, *country*, *isle*, *forest*, *landmark*, *mountain*, *state*, and *waterbody*.

provide citizens in their vicinity with basic facilities and services such as schools, post offices, banks, and supermarkets. Middle centers have a larger population and range, and typically offer facilities such as hospitals, cinemas, secondary schools, lawyers, and medical specialists. Places are referred to as upper centers if they offer facilities such as specialized stores and clinics, museums, universities, and regional administration offices. As such, the list of places categorized into these classes provides a useful reference dataset to develop and test our extraction approach.

The analysis is limited to upper and middle centers here for several reasons. In many states, the declaration of lower centers is only done at the regional planning level (Bayerisches Staatsministerium der Finanzen, für Landesentwicklung und Heimat, 2016, for example). Since these plans are not available in machine-readable formats, they would require a manual extraction for each of those regional plans (Bavaria alone has 18 regions). After extraction, they would have to be manually matched against their corresponding Wikipedia pages, bearing in mind that there are often multiple villages and towns that go by the same name across Germany. The fact that just North Rhine-Westphalia already has 189 lower centers (Staatskanzlei des Landes Nordrhein-Westfalen, 2015) gives an impression of the amount of manual work required to accomplish this task for all of Germany. Moreover, it is questionable whether the resulting dataset would be useful for our study at all. Lower centers are usually relatively small places, so that their representation in Wikipedia is varying significantly both in terms of article extent and number of incoming links. Finally, Berry and Garrison (1958a) already found that some of the principles of CPT do not seem to apply anymore once the population density or levels of urbanization come below a certain level. For those reasons, the analysis presented here is limited to upper and middle centers.

Among the 16 German states, Berlin, Hamburg and Bremen have a special status. The city states Berlin and Hamburg only consist of a single city, whereas Bremen also contains Bremerhaven as enclave surrounded by Lower Saxony. In these three states, the zoning plan replaces the state development plan, and hence no central places are being defined. As these places clearly do play a central role for their vicinity, they have been classified as upper centers (Berlin, Hamburg, Bremen) and middle centers (Bremerhaven), respectively, for this study. With those changes, the reference dataset contains the 123 upper centers and 874 middle centers shown in Figure 3. Some of them are part of a group of cities or towns that fulfill the functions of a middle center in combination.[3]

3.3 Data Preparation

Two datasets were downloaded for this study from the Wikipedia dump archive, one containing a full dump of the articles in the German Wikipedia, the other one containing all geotagged pages.[4] The articles were then parsed to extract all pages by title, as well as links and mentions between any two pages. In a subsequent step, the total number of incoming links and mentions was calculated

[3]The exact German categories for middle centers considered here are *Mittelzentrum*, *Mittelzentrum mit Teilfunktion eines Oberzentrums*, *Mittelzentrum im Verdichtungsraum*, and *Mittelzentraler Verbund*.

[4]Dumps for the German language Wikipedia are available from `https://dumps.wikimedia.org/dewiki/`.

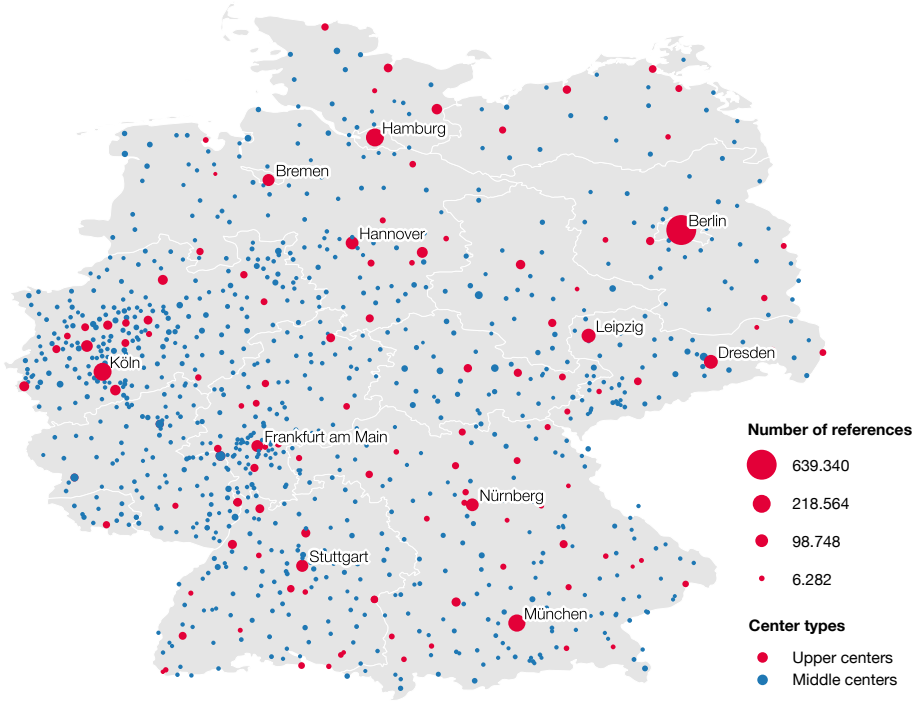

Figure 3: Map of upper and middle centers in Germany. Symbol size indicates the number of incoming references for their respective Wikipedia pages.

for every page. The geotagged pages dump was used to extract the geographic coordinates for all pages that contain such geotags, along with its classification. Finally, the geographic distance between the geotag coordinates was calculated for any two pages that have (a) at least one link between them, and (b) both of them are geotagged. Spherical distance was used to speed up the distance calculation for all 3 million links between geotagged pages. The distortion introduced by this simplification should make no difference at the relatively local scale of this study.

The upper and middle centers in the reference dataset have been extracted manually from Wikipedia[5] and subsequently checked for any linking errors, i.e., it was checked that every place listed has a Wikipedia page with a geotag. Some errors had to be fixed manually, most of which linked to the wrong Wikipedia page. The scripts that implement this process are available online for inspection and reuse, along with detailed instructions for using them.[6] Besides the parsing process described above, their main purpose is to get the data into a PostGIS database that supports efficient querying and spatial analysis. Figure 4 shows a conceptual overview of the database generated by this process.

[5]See `https://de.wikipedia.org/wiki/Liste_der_Ober-_und_Mittelzentren_in_Hessen` for the upper and middle centers in Hessen, for example.

[6]See `https://github.com/crstn/CentralPlaceWiki`.

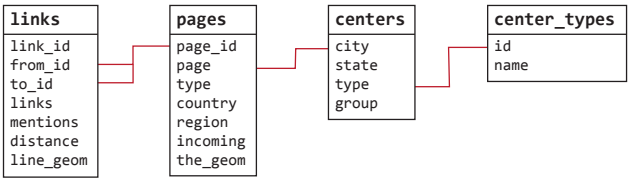

Figure 4: Overview of tables and relationships in the PostGIS database generated for this study.

4 The Characteristics of Central Places in Wikipedia

This section present different aspects of central places reflected in their corresponding Wikipedia pages. It introduces an exploratory analysis tool and provides quantitative analyses of the link structure between their pages, as well as the spatial configuration of the central place hierarchy.

4.1 Exploratory Analysis

In order to be able to quickly test and visualize different approaches to extract the centers in the vicinity of a place, a browser-based application has been developed that connects to the PostGIS database and visualizes the results of a place query on top of a web map.[7] When the user selects a place name from the search results, its predicted network of lower-level places is calculated. Each place in that network becomes a node, which can be clicked to load its own network, which is rendered in a different color.

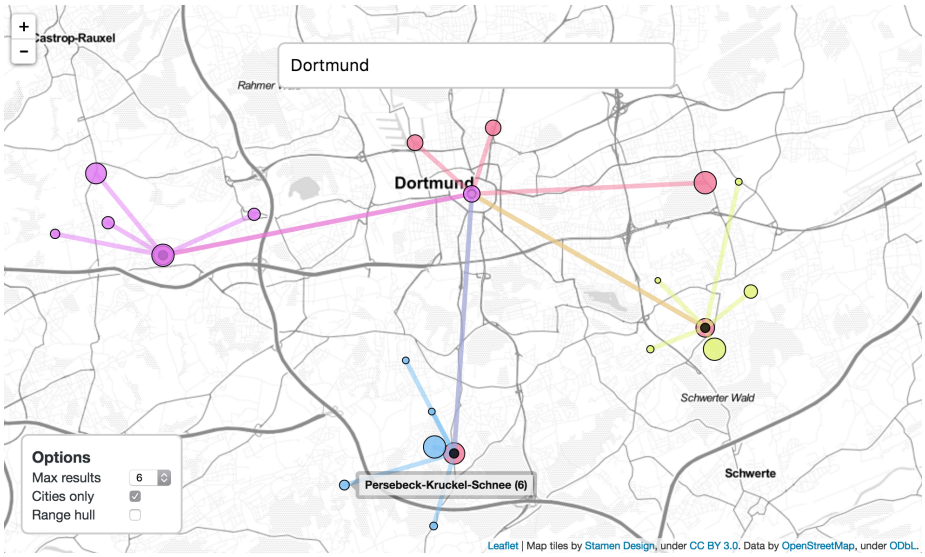

Figure 5: Screen shot of the browser-based application for exploratory analysis.

[7] The source code for the tool is available from `https://github.com/crstn/CentralPlaceWiki/tree/master/webapp`.

This tool has proven extremely useful when testing different extraction approaches, as it allows for a quick modification of SQL queries in the back end of the tool. The effect of these changes on the results can then be quickly explored visually for a large number of places. The same process would be very cumbersome through a desktop GIS, which would require re-writing and manually executing the SQL query for every individual place.

Exploring the dataset this way provided three main insights. First, it is not possible to meaningfully extract central place structures without knowing which pages are about cities. If the corresponding option in the tool is turned off, many of the results returned are pages about airports or universities, for example. While it makes sense that they refer frequently to the city they are located at, they are not relevant in the context of the task at hand. This confirms that some semantic knowledge is required for this task (Keßler, 2015). Second, it shows that the coarse classification of pages in Wikipedia is a challenge for the extraction of a central place structure, as many neighborhoods or districts are also classified as *city*, even though they are only parts of a city in reality. Third, it showed that the geotags in Wikipedia need to be taken with a "grain of salt", not just in terms of precision (which is not exactly a new insight; see Janowicz et al. (2016)). Some of the geotags located the places in completely wrong locations, such as several of Hamburg's districts that have been moved to the West coast of Ireland.

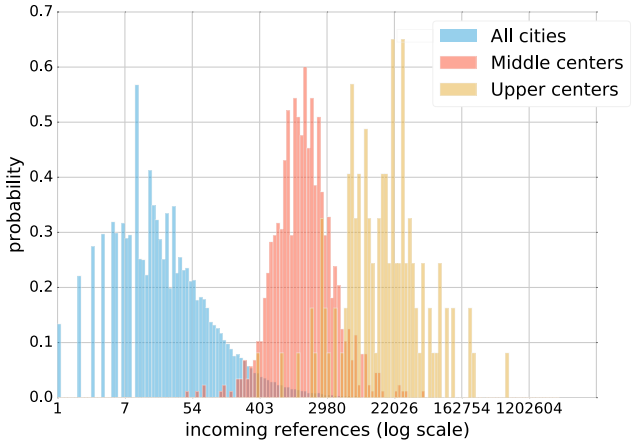

Figure 6: Normalized histogram of incoming references for upper and middle centers; data for all pages classified as *city* shown for reference. Note that the bins are plotted on a logarithmic scale, i.e., the numbers for incoming links for middle and upper center are significantly larger than for cities overall.

4.2 Link Frequency

Following the initial exploratory analysis, a more systematic analysis of the characteristics of the centers in the reference dataset has been conducted. The map shown in Figure 3 indicates that most upper centers get more incoming references than the middle centers. Figure 6 confirms this impression, however, it also shows that there is a significant overlap in the bins between middle centers

and upper centers, i.e., it is not possible to find out whether a city is an upper or middle center solely based on the number of references to its Wikipedia page. Looking at this problem from an information retrieval point of view confirms that the simple approach of identifying centers by number of incoming references works for most cases, but not for all. If we take the 123 german cities with the highest number of incoming references to their Wikipedia pages (since there are 123 upper centers), 89 of the actual upper centers are among the results. This corresponds to a recall of 0.725. The 250 most referenced city pages already contain 90% of the upper centers (recall 0.9, precision 0.4). However, to retrieve all 123 upper centers – i.e., to reach a recall of 1.0 – we have to get the 3361 most referenced pages of cities. The precision in this case drops to under 0.04.

A notable insight from this study is that the number of references to the Wikipedia page of a place is almost a perfect correlation with its population number (Pearson's $\rho = 0.96$, $p < 0.001$), as shown in Figure 7. The same figure also shows that some places are declared centers despite very low population numbers and number of incoming references, which renders an automatic classification challenging.

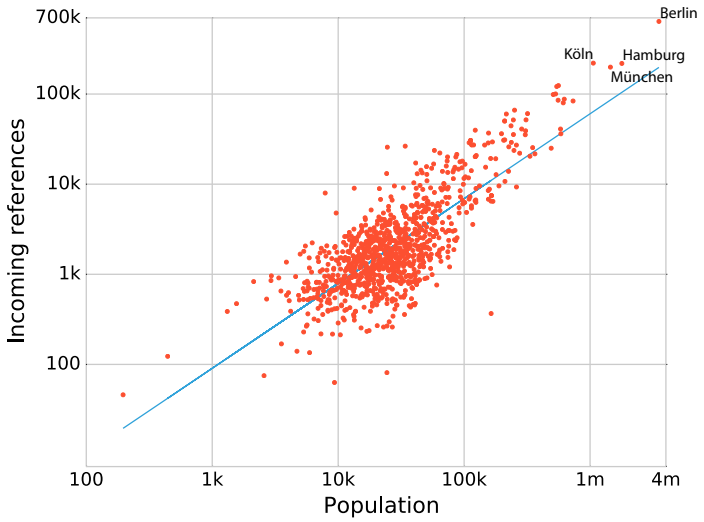

Figure 7: Log-log scatter pot of incoming references of all centers compared to the population of the corresponding place.

4.3 Spatial Aspects

The spatial aspects of the link structure of a city's Wikipedia page can be analyzed by looking at the distribution of geographic distances to the places whose pages are linking to it. Figure 8 shows a boxplot of distances to those places for incoming links to upper centers, middle centers, and all other cities. Upper centers have a significantly larger "link range" than middle centers, i.e., they generally also receive more incoming links from pages about places that are further away. For upper centers, the median distance is 40.0km (25% quartile at 15.5km, 75% quartile at 95.7km), whereas the median distance for middle centers is 13.0km (25% quartile at 6.4km, 75% quartile at 31.2km). Likewise,

middle centers have a larger link range than cities that are no centers of any kind; the difference is smaller, however, with the median distance for all other cities at 8.8km (25% quartile at 3.2km, 75% quartile at 23.2km).

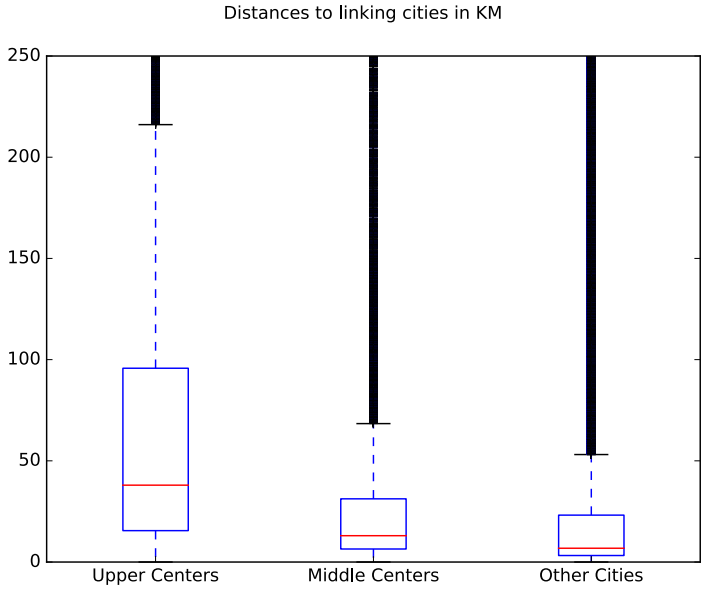

Figure 8: Boxplot of distances for incoming links to upper centers, middle centers, and all other cities. The plot has been capped at 250km distance, outliers in all three categories go up to several thousand km.

While it may seem that pages about nearby cities generally link to each other more often, a statistical evaluation does not support this assumption. There is no correlation between the number of mentions a page receives from another one and the distance between the cities the pages are about. An analysis of the link structure between the pages about the middle and upper centers in the reference dataset, however, reveals a significant "traction" towards a nearby upper center. Figure 9 shows the same map as Figure 3, however, it adds a link to the most-referenced center for every middle center. For most middle centers, the most-referenced center is a nearby upper center. While it is often the closest one, cities such as Berlin or Hamburg also appear as most-referenced centers for cities that do have another closer upper center. Some middle centers also form networks without a central upper center, such as in the North-Western corner of the country. It is worth noting that some centers are disconnected from this network. There are upper centers whose pages do not show as the most-referenced page of any middle centers (isolated red dots). Likewise, some middle centers are disconnected from the network whose pages are not the most-referenced page of any middle centers, and, more interestingly, also do not have outgoing references to *any* of the other centers (isolated blue dots). Overall, there are 10 disconnected middle centers out of 700 (1.4%), and 23 disconnected upper centers out of 123 (18.7%). All of those 23 upper centers are actually the most referenced pages from pages about other upper centers, however, which shows that there is also a strong inter-regional network connecting the upper

centers.

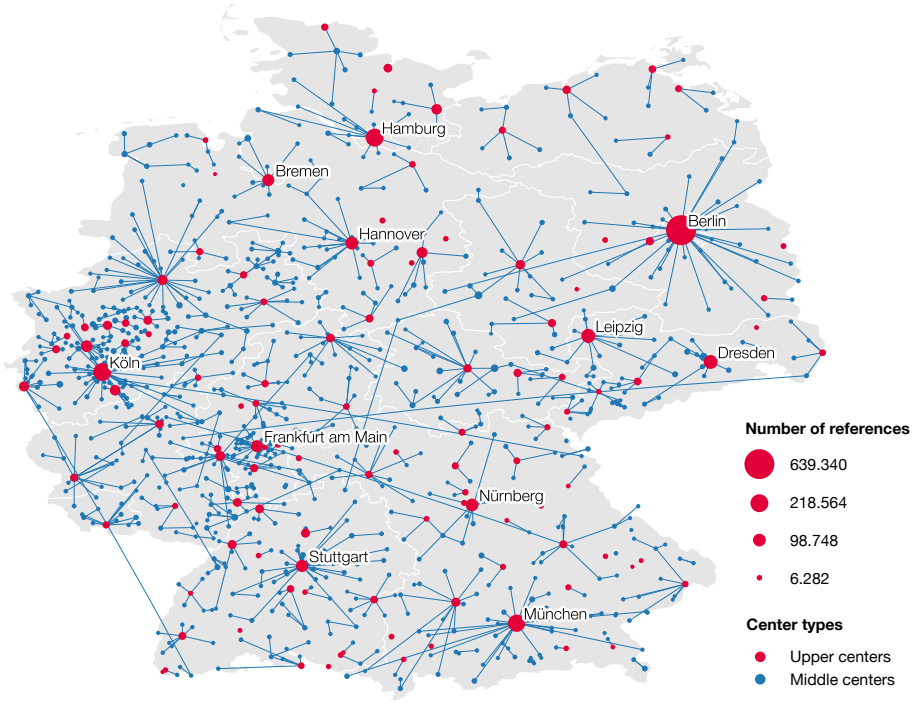

Figure 9: The same map of upper and middle centers as in Figure 3, but with links showing the most-referenced center for every middle center.

5 Extracting Central Places from Wikipedia's Link Structure

This section discusses a top-down and a bottom-up approach towards the extraction of central places from Wikipedia and analyzes the results.

5.1 Top-down vs. Bottom-up Extraction

Using the insights gained in Section 4, different approaches have been tested to extract a central place structure from Wikipedia, solely based on the link structure between the corresponding pages. An initial attempt followed a *top-down* approach, starting with the fact that the largest regional centers have the pages with the highest number of incoming references. The next level of centers was then hypothesized to consist of the pages with the highest number of references to those regional center pages. This attempt has not proven useful because of the very coarse classification system, which also includes many city districts and neighborhoods classified as *city* (see Section 3.1). These city districts and neighborhoods produce the highest number of references for a vast majority of the upper centers, so that the top-down approach does not produce any useful results.

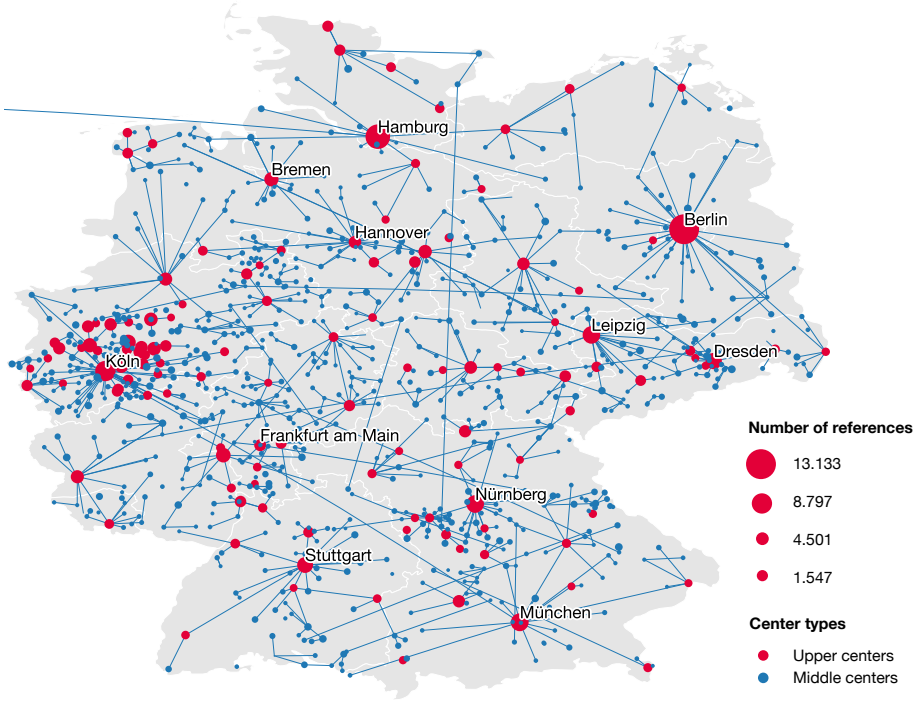

Figure 10: Network of central places extracted from Wikipedia link structure.

A *bottom-up* approach has proven more useful instead. For every city in the dataset, the most-referenced city page has been collected (similar to the way the links between places in Figure 9 have been produced). From this survey that provides the most-referenced city for *every* city in Germany (hence bottom-up), the total number of *incoming* references has been summed up to produce a ranking of most-referenced cities. The centers are then based on the top 1000 results in this ranking, assigning the top 125 results upper center status, and middle center status to the rest. This corresponds roughly to the total number of official centers and percentage of upper centers among them in Germany.

5.2 Comparison of Extracted and Official Centers

The result of the bottom-up extraction is shown in Figure 10, with significant similarities to the network of official upper and middle centers shown in Figure 9. The largest regional centers are all present in the map produced bottom-up, often with a large overlap with the official centers in the network of middle centers around them. Concerning the differences between the two networks, the bottom-up network seems to predict more upper centers in densely populated areas, such as in the Ruhr area North of Cologne (*Köln*). Moreover, the network appears to be more "messy", which goes back to the larger number of long-distance links. These can often be found between twin towns and sister cities in different parts of the country. A systematic evaluation of this phenomenon is hard to conduct because a central data source listing all such partnerships between cities does not seem to exist. A manual check of some of the long-

distance links between centers confirmed this impression.

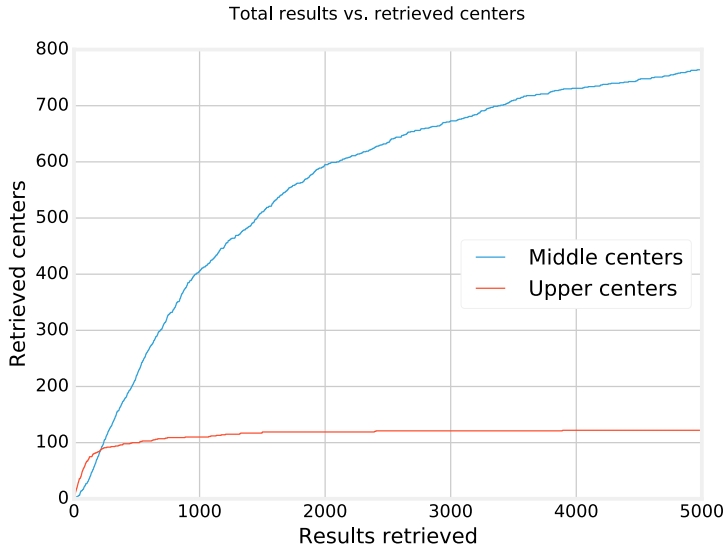

Figure 11: Number of retrieved upper and middle centers declared in Germany that could be extracted compared to the total number of retrieved results.

In order to quantify the degree of overlap between the bottom-up place hierarchy based on Wikipedia and the official place hierarchy, an analysis of the upper and middle centers retrieved has been conducted based on the method described above. Figure 11 shows the number of actual upper and middle centers retrieved from Wikipedia's link structure per total number of results retrieved (the map in Figure 10 shows 1000 results). Most upper centers are among the first results, but the smaller upper centers only appear among the results when the list is expanded significantly. Likewise, many middle centers are only retrieved when the total number of results reaches several thousand. This confirms the big overlap between many middle centers and other cities in the country (see Figure 6), which makes them hard to distinguish. Nevertheless, 110 out of 123 official upper centers and 404 out of 874 official middle centers could be extracted in the first 1000 results using the bottom-up method (recall and precision of ∼0.51).

5.3 Spatial analysis

In addition to the analysis focusing on retrieved centers, a spatial analysis has been conducted in order to understand how the extracted network of centers differs from the official central place structure. For this purpose, both sets of centers have been considered as point patterns. For every officially declared center, the nearest neighbor in the point pattern with the extracted centers has been identified[8] Figure 12 shows the results of this analysis for the state of North Rhine-Westphalia.

[8]This analysis as been conducted using the `nncross` function in the `spatstat` package for R (Baddeley et al., 2015).

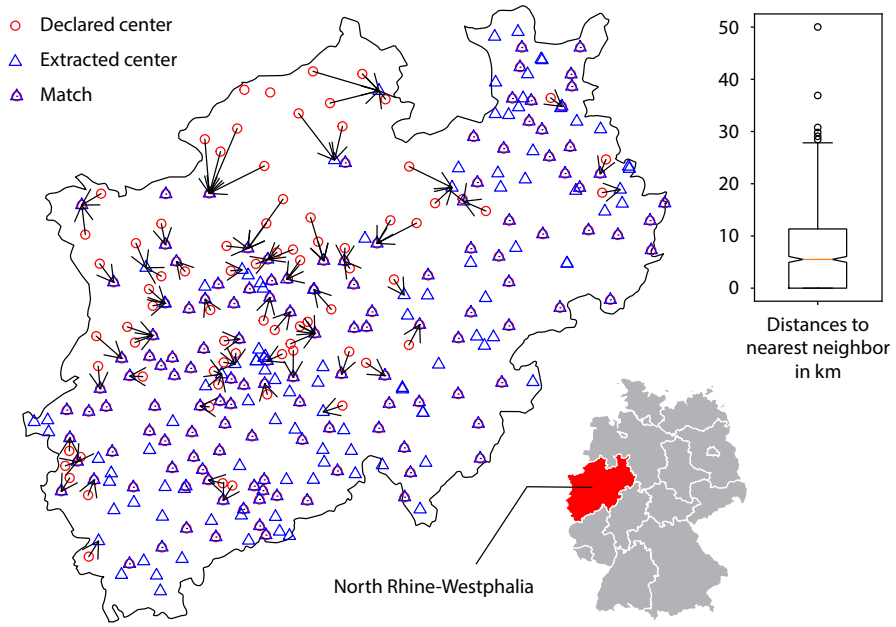

Figure 12: Map of nearest neighbor in the extracted centers layer for every officially declared center (left) and box plot of distances to nearest neighbors across all of Germany (right).

The mean distance to the nearest neighbor is 6.1km (median: 0km; standard deviation: 8.3km), which points to a high similarity of the two point patterns. This is not surprising, since 511 centers (51%) can be found in both layers. In order to test whether the two patterns differ in clustering behavior, they have been analyzing using Ripley's K-function using the outline of Germany as a window. The results also showed no significant differences. Finally, the density of declared and extracted centers has been compared. Figure 13 shows that both patterns have areas with high density cells in the West of the country, and areas with low density cells in the North-East. Besides these general trends, there are several smaller areas and individual cells that show significant differences. These differences result from cases where the extraction approach selects a range of cities and towns in one area, when the declared centers are actually in a neighboring area. Such a case is also shown in Figure 12, where most of the declared centers are in the North-West of the state, whereas the extraction algorithm selects more cities as centers in the South-East. It is unlikely that such differences can be resolved with an approach solely based on links and mentions.

6 Conclusions

The research presented in this paper focused on analyzing the link structure between Wikipedia pages about upper and middle centers in the central place hierarchy of Germany. It has found that the significance of a center is reflected

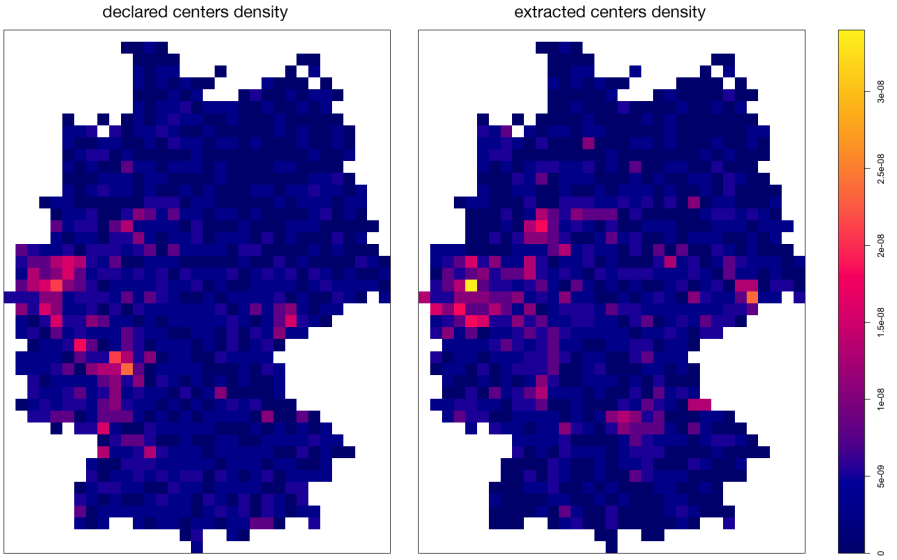

Figure 13: Comparison of density of centers at 20km resolution.

in the number of incoming references to its Wikipedia page and in a larger link range, i.e., large regional centers not only attract more references, they also come from pages about places that are further away. These insights have then been used to develop a bottom-up method that extracts central places from Wikipedia solely based on the references between the places' Wikipedia pages. This bottom-up place hierarchy shows significant overlap with the official place hierarchy, especially concerning the upper centers. While there is less overlap at the level of middle centers, this result can be explained with the large similarities in link structure between middle centers and cities that do not function as a center in the official place hierarchy. It is also intuitive given the large number of middle centers declared for Germany. A limitation of the bottom-up approach is the fact that it relies on the Wikipedia classification to find pages about cities. While it does not seem feasible to be able to tell apart pages about cities from other Wikipedia pages based on the link structure alone, future research could look into text mining techniques to automate this process (Nakayama et al., 2008) and potentially also identify missing links between cities. This would also present an opportunity for a more fine-grained classification of neighborhoods and districts, which are currently all classified as cities in Wikipedia. A further limitation of using the number of mentions of a page as a measure of its relative relevance is the potential confusion of place names with common terms. While this was not a problem for Germany, one could easily imagine that this might become a problem when using the same approach in other languages. A similar problem did arise for places with similar names, though. Figure 10 shows a long-distance link between Munich (München) in the South and a place near Cologne in the West. It turns out that this place is Mönchengladbach, which used to be called München-Gladbach until 1950. The name change was made in order to avoid confusion with München – this fact is explained on Mönchengladbach's wikipedia page with a link to the page about

München. Somewhat ironically, this lets the extraction algorithm believe that any mentions of München-Gladbach actually refer to München. While I am not aware of any other concrete examples of such confusions, I am certain that more do exist in the dataset, but with less prominent effects.

In the larger context of Volunteered Geographic Information, the results of this research are a first indicator that real-world relationships between geographic entities are indeed reflected in the structure of the web. This is comparable to the more obvious fact that online social networks reflect real-world interpersonal relationships: Social networks not only contain information about the individual users, but also explicit (being connected to someone) and implicit (often peeking at someone else's profile, for example) information about their relationships. Likewise, there is a plethora of explicit information on the web about individual places (such as Wikipedia pages or gazetteer entries) and relationships between them (such as their administrative hierarchy), but there also seems to be an additional layer of implicit relationships to be revealed. In order to confirm that this is a broader effect rather than a peculiarity of the German language edition of Wikipedia, more research is required to confirm the findings of this paper. An obvious first choice would be other language editions of Wikipedia to confirm the findings of this research also for regions outside of Germany. Potential follow-up studies should also look at other kinds of user-generated content. Does the network of Twitter users following each other reveal any structural information about their home locations? Is it possible to derive the spatial context of an article on the New York Times website based on an analysis of the websites linking to that article? Can the locations of viewers and likers of a YouTube video tell us something about the places covered in the video? Can machine learning techniques be trained with ground truth data, such as the reference datasets used in this study, to automatically identify distinguishing features of upper and middle centers? Answering these questions is by no means straightforward, nor is getting access to the required data. Working on these challenges, however, will provide us with a better understanding of how big the influence of geography really is in a medium that was once imagined to make location irrelevant.

References

Adams, B. and McKenzie, G. (2012). Frankenplace: An Application for Similarity-Based Place Search. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*.

Baddeley, A., Rubak, E., and Turner, R. (2015). *Spatial Point Patterns: Methodology and Applications with R*. Chapman and Hall/CRC Press, London.

Baskin, C. W. (1966). *Central Places in Southern Germany*. Prentice Hall.

Bayerisches Staatsministerium der Finanzen, für Landesentwicklung und Heimat (2016). Regionalpläne. Available from `https://www.landesentwicklung-bayern.de/instrumente/regionalplaene/`.

Berry, B. J. L. and Garrison, W. L. (1958a). A Note on Central Place Theory and the Range of a Good. *Economic Geography*, 34(4):304–311.

Berry, B. J. L. and Garrison, W. L. (1958b). The Functional Bases of the Central Place Hierarchy. *Economic Geography*, 34(2):145–154.

Brush, J. E. (1953). The Hierarchy of Central Places in Southwestern Wisconsin. *Geographical Review*, 43(3):380–402.

Christaller, W. (1933). *Die zentralen Orte in Süddeutschland*. Gustav Fischer, Jena.

Gao, S., Li, L., Li, W., Janowicz, K., and Zhang, Y. (2017). Constructing gazetteers from volunteered Big Geo-Data based on Hadoop. *Computers, Environment and Urban Systems*, 61, Part B:172–186.

Goodchild, M. (2007). Citizens as Sensors: The World of Volunteered Geography. *GeoJournal*, 69(4):211–221.

Hardy, D. (2010). Volunteered geographic information in Wikipedia. PhD thesis, University of California, Santa Barbara.

Hecht, B. and Moxley, E. (2009). Terabytes of Tobler: Evaluating the First Law in a Massive, Domain-Neutral Representation of World Knowledge. In Hornsby, K., Claramunt, C., Denis, M., and Ligozat, G., editors, *Spatial Information Theory*, volume 5756 of *Lecture Notes in Computer Science*, pages 88–105. Springer Berlin Heidelberg.

Hecht, B. J. and Gergle, D. (2010). On the localness of user-generated content. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 229–232.

Hsu, W.-T. (2012). Central Place Theory and City Size Distribution. *The Economic Journal*, 122(563):903–932.

Janowicz, K., Hu, Y., McKenzie, G., Gao, S., Regalia, B., Mai, G., Zhu, R., Adams, B., and Taylor, K. (2016). Moon landing or safari? a study of systematic errors and their causes in geographic linked data. In Miller, J. A., O'Sullivan, D., and Wiegand, N., editors, *Geographic Information Science: 9th International Conference, GIScience 2016, Montreal, QC, Canada, September 27–30, 2016, Proceedings*, pages 275–290, Cham. Springer International Publishing.

Keßler, C. (2015). Central Places in Wikipedia. In Bacao, F., Santos, Y. M., and Painho, M., editors, *AGILE 2015: Geographic Information Science as an Enabler of Smarter Cities and Communities*, pages 35–52. Springer International Publishing.

Keßler, C., Janowicz, K., and Bishr, M. (2009). An agenda for the next generation gazetteer: Geographic information contribution and retrieval. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '09, pages 91–100, New York, NY, USA. ACM.

Keßler, C., Maué, P., Heuer, J. T., and Bartoschek, T. (2009). Bottom-Up Gazetteers: Learning from the Implicit Semantics of Geotags. In Janowicz, K., Raubal, M., and Levashkin, S., editors, *Third International Conference*

on GeoSpatial Semantics (GeoS 2009), Springer Lecture Notes in Computer Science 5892, pages 83–102.

Lieberman, M. D. and Lin, J. (2009). You Are Where You Edit: Locating Wikipedia Contributors through Edit Histories. In *Proceedings of the Third International Conference on Weblogs and Social Media, ICWSM 2009, San Jose, California, USA, May 17–20, 2009*.

Lösch, A. (1954). *The Economics of Location*. Yale University Press, New Haven, CT.

Nakayama, K., Hara, T., and Nishio, S. (2008). Wikipedia link structure and text mining for semantic relation extraction. In Bloehdorn, S., Grobelnik, M., Mika, P., and Douc, T. T., editors, *Proceedings of the Workshop on Semantic Search (SemSearch 2008) at the 5th European Semantic Web Conference (ESWC 2008), Tenerife, Spain, June 2nd, 2008*, pages 59–73.

Openshaw, S. and Veneris, Y. (2003). Numerical experiments with central place theory and spatial interaction modelling. *Environment and Planning A*, 35(8):1389–1404.

Overell, S. and Rüger, S. (2006). Identifying and grounding descriptions of places. In *SIGIR Workshop on Geographic Information Retrieval*, pages 14–16.

Raumordnungsgesetz (2015). Raumordnungsgesetz vom 22. Dezember 2008 (BGBl. I S. 2986), das zuletzt durch Artikel 124 der Verordnung vom 31. August 2015 (BGBl. I S. 1474) geändert worden ist. Available online from `https://www.gesetze-im-internet.de/bundesrecht/rog_2008/gesamt.pdf`.

Salvini, M. M. (2012). Spatialization von nutzergenerierten Inhalten für die explorative Analyse des globalen Städtenetzes. PhD thesis, University of Zurich. Available from `http://www.zora.uzh.ch/72166/`.

Santos, D. and Cardoso, N. (2008). Gikip: Evaluating geographical answers from wikipedia. In *Proceedings of the 2nd international workshop on Geographic information retrieval*, pages 59–60. ACM.

Staatskanzlei des Landes Nordrhein-Westfalen (2015). Landesentwicklungsplan Nordrhein-Westfalen. Überarbeiteter Entwurf Stand 22.05.2015. Available from `https://www.land.nrw/sites/default/files/asset/document/01_10_2015_lep_text_zweite_beteiligung_lanuv.pdf`.

Takahashi, Y., Ohshima, H., Yamamoto, M., Iwasaki, H., Oyama, S., and Tanaka, K. (2011). Evaluating significance of historical entities based on tempo-spatial impacts analysis using wikipedia link structure. In *Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia*, pages 83–92.

Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, 46:234–240.