Geolocalization of digital data

Daniela Stoltenberg, Barbara Pfetsch, and Annie Waldherr

When digital technologies first began to proliferate, some predicted the "death of distance" (Cairncross 1997). By now, however, we have come to realize that location and place have by no means lost significance in or as a result of digital media. Location information is embedded in digital data in many ways. For example, social media platforms allow users to disclose where they live in their profiles. Blog posts, news stories, and comments discuss places, and platforms like *Twitter* and *Instagram* make it possible to display the location where a post is made (see Lettkemann in this volume).

These applications have significantly increased the amount of data researchers can use to study spatial contexts in the digital realm. At the same time, the characteristics of this "big data" (Russom 2011: 6)—namely *volume* (size and required storage space), *velocity* (highly dynamic, undergoing change in real time), and *variety* (differing formats)—make the traditional methods of the social sciences largely inapplicable. Automated methods have become necessary.

These data and methods open up numerous innovative applications, triangulation opportunities, and interesting interpretations of content for qualitative spatial research. For example, geoinformation can be used to identify user groups whose interaction with spatial information can then be investigated in ethnographic studies or qualitative interviews: Why do some users share their locations in social networks, while others choose not to disclose location information? Geoinformation in digital data can also be a prerequisite for the qualitative analysis of local networks of people or groups. The reconstruction of user types based on digital location information can then, for example, be combined with qualitative network analyses (Hollstein/Straus 2006) to describe spatial and social interaction patterns. In discourse analysis, the frequency of references to specific locations in unmanageably large corpora of text can be important in describing patterns of meaning. If we can quantify *that* some locations are being discussed in a notable way, we can elaborate on this observation with qualitative and mixed-methods discourse analysis (Duchastel/Laberge 2019).

Finally, the automated analysis of location information in digital data opens up entirely new possibilities for the spatial contextualization of qualitative descriptions of social behavior. If information about the interactive space of people or groups has been obtained in a qualitative network analysis, this can then be combined with data about

their online communication to ask: What is the spatial and social structure of personal networks and how are these subjectively perceived? Or, for a location-specific discourse analysis: What other places are also evoked in the relevant communications? How do discourses about specific locations differ?

Even qualitative spatial researchers should view the automatic identification and classification of georeferences in large corpora of text without blinders and exploit these methods in their research. As the examples above show, the triangulation and iterative combination of digital big data applications and qualitative ethnographic analytical techniques open up entirely new horizons of research. At the same time, automatic procedures for extracting location information from texts are becoming more and more accessible even to researchers without a computing background and corresponding research designs are becoming easier to implement.

Given this background, this chapter aims to present simple automatic geocoding methods that, when combined with qualitative methods, facilitate original perspectives in spatial analysis. We begin with the characteristics of location information in digital datasets and the challenges posed by their analysis. We then give an overview of approaches to the automatic classification of location information and discuss their strengths and weaknesses. We introduce the practice of automatic analysis of digital geoinformation in two examples from our own research.[1] Finally, we take a step back and problematize questions of scientific ethics.

1 Properties of digital location information

References to place appear in digital media in many forms. As they are authentic manifestations of human behavior, they are considered observational data. Although geodata is easy to collect for scientific analysis, they are rarely standardized in something like a coordinate pair (e.g., GPS data from smartphones). Usually, unstructured location information poses significant challenges for researchers. This is illustrated by georeferences on the microblogging platform *Twitter*. Three forms of references to place are particularly relevant: (1) *geotags* in individual messages (tweets), (2) *location information in profiles* of users, and (3) mentions of places in the *text of tweets* (see Wilken 2014).

Geotags are the form of location information in *Twitter* data that is easiest to process. They provide information about the location of users at the time of tweeting. The location information is attached to the tweet as metadata and is generated by the GPS system of mobile devices. Geotags are accessible through *Twitter's* programming interface (*Application Programming Interface*, API). The name of the location, a description of the geographic unit (e.g., city), and degrees latitude and longitude are given as separate variables. However, geotag analyses can distort results considerably: Only about one percent of users around the world use the geotagging function, and this population differs sociodemographically from the general user base (Malik et al. 2015).

1 The research project *Translocal Networks: Public Sphere in the Social Web* of Collaborative Research Centre 1265, "*Re-Figuration of Spaces*," funded by the German Research Foundation (DFG).

The information in the *"location" field of user profiles* provides much better coverage. Depending on the user sample, roughly 80 percent of all profiles have text in this field (Hecht et al. 2011: 240; Kinsella et al. 2011: 64). It is a free text field where users can input their place of residence. Unlike a geotag, the location is manually entered information. Consequently, the location field may not necessarily contain useful geoinformation: Users can input entries that do not refer to real places. Hecht et al. (2011: 240) concluded that, in a sample of 10,000 US American profiles, a total of 66 percent contained references to real places. On the other hand, 18 percent of users entered no information and 16 percent entered non-geographical information. Problems can also occur because of spelling errors and the diversity of possible languages. Since the location field is a free text field, geographical information can be entered to any degree of detail, from an exact mailing address to the planet Earth. In addition, users can enter more than one place into the text field, causing problems for automatic detection. In Hecht et al. (2011: 241), this applied to 2.6 percent of all profiles.

Finally, *references to places in the body of tweets* can also be of interest. Unlike in the case of geotags and location fields, these are not so much the places where users reside, but rather those they are talking about. Such references pose the greatest challenge for automatic detection as, unlike geotags and location fields, it is unclear where to search for location information. In addition, the fraction of relevant references in the total volume of text is usually very low, increasing the risk of false positive coding, that is to say, extracting references to places where there are none.

Once such references have been found, researchers must consider how informative and high-quality the data is. We know that a substantial portion of users in most online datasets do not provide explicit information about their whereabouts. Assuming that these users may also differ from those who do share their whereabouts (e.g., sociodemographically, in their networking or communication behavior), this has consequences for the explanatory power of studies on socio-spatial behavior based on this data.

In summary, the problems of big data analysis (Russom 2011: 6) also apply to spatially determined digital (*Twitter*) data. These datasets are characterized above all by their sheer size (*volume*). They are too big for researchers to obtain a general overview of the embedded location information through manual coding. Secondly, the location information has considerable variance (*variety*): It varies in its structure, degree of detail, language, correctness, etc. Thirdly, the datasets themselves are undergoing continuous change (*velocity*) as new communications are constantly being added. In addition, the structure of the data can change because of the specific platform architectures.

2 Approaches to quantitative classification

The classification of geoinformation in digital data can have different goals, such as standardization of data according to a predetermined format (e.g., "city, country"), annotation with coordinates, or determination of geographical level. This means there is no standard procedure or best practice (Hoffmann/Heft 2020: 7) but rather an overwhelming number of different approaches and tools. Fundamentally, these fall into two categories: procedures that identify and standardize mentions of location and procedures

that attempt to infer missing geoinformation. Each category contains a variety of approaches (see Fig. 1).

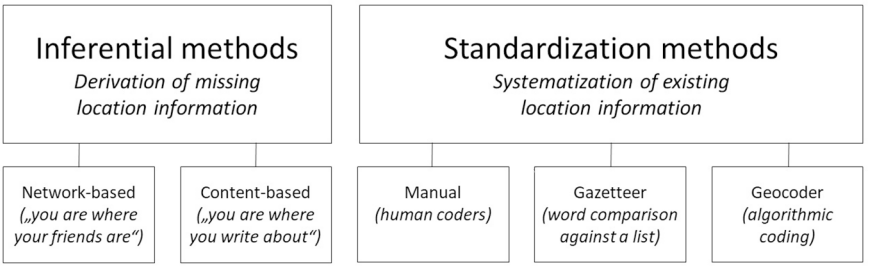

Fig. 1: Methods for extracting location information from digital data | ©Author's own diagram

Inferential methods can be divided into network-based and content-based methods. Network-based methods are based on the principle "you are where your friends are" (Rout et al. 2013), i.e., on the observation that social relationships are predominantly formed with people nearby. They use connections with users who have explicit locations to infer missing locations (see McGee et al. 2013; Sadilek et al. 2012). Content-based methods, on the other hand, are based on the principle "you are where you write about" (Rout et al. 2013): the assumption that users tend to generate content specific to their location. Machine learning approaches are useful here as they can identify topics strongly associated with particular locations (see Gore et al. 2015; Kinsella et al. 2011).

These inferential methods depend on many assumptions and are fraught with numerous problems (Hoffmann/Heft 2020: 7). Often only about 50 percent of their attributions are valid, or attribution is only possible to very rough territories. Their performance is also difficult to test, as a *ground truth* only exists for users who also provide their whereabouts.

Standardization of available geoinformation can be subdivided into three approaches: manual coding, gazetteers (geographical dictionaries), and automatic geocoding service providers. Manual coding can be considered the gold standard (Takhteyev et al. 2012: 76; Hoffmann/Heft 2020: 11). Human coders have a high level of contextual knowledge, allowing them to make complex assessments. A human coder will generally identify "Paris" as the French capital and will usually be right. In contrast, considerable training data and rules are required to reliably prevent automatic methods from choosing one of the 17 US American cities with the same name (Hoffmann/Heft 2020: 5). Human coders can also situationally use appropriate sources (e.g., maps, encyclopedia) to make decisions and can also consider when the available information does not allow for coding. With respect to quality criteria in the social sciences, the use of human coders maximizes validity. However, manual coding is time-consuming and expensive. The work is also tiring and error-prone, compromising reliability. For this reason, manual coding is a valid approach for small datasets, but possesses limited scalability. In contrast, computer-

assisted methods like gazetteers or algorithmic geocoders can be applied to arbitrarily large datasets.

A (digital) gazetteer (geographical dictionary) contains structured information about geographical places. A location can have multiple names, and locations can be identified at every geographical level (e.g., neighborhood or city). These need not be officially defined units; informal names can also be added (Goodchild/Hill 2008: 1041).[2] The location names contained in a gazetteer can be compared to the text to find references to places. This approach consists of dictionary-based location coding in which the constructs of interest (here: locations) are defined in advance and associated with corresponding features (here: location identifiers). These features are then compared to the words (tokens) in the texts (Scharkow 2013: 299). As the method uses word comparison, it is perfectly reliable. It will always return the same result if applied multiple times to the same study material. Validity, on the other hand, is only as good as the effort invested in constructing the dictionary. There are also limitations with respect to granularity and/or scalability. If the data refers to a limited geographic region, it is possible to achieve high coverage even of smaller places. If one attempts to scale to the global level, however, the problem of ambiguous names becomes acute.

One efficient way to code large amounts of location information is "geocoders": algorithmic applications that identify, standardize, and annotate location information in textual data. Geocoders are typically commercial services such as *Google Maps*. Depending on the use case, other providers with flat rate pricing (e.g., *OpenCage*) or even free services (e.g., *Datascience Toolkit Geocoder*) are possible alternatives. These applications are normally accessible through an API, to which researchers can automatically submit large numbers of queries. The process is highly efficient and requires very little programming skill. The entered input (e.g., the entry in a location field) is algorithmically processed and compared to databases (e.g., belonging to *Google Maps* or *Open Street Map*). As these databases have a high coverage of geographical information at all levels, a city is highly likely to be coded as accurately as a country or an exact address. Geocoders are characterized by considerably greater flexibility with respect to the input than dictionary-based approaches. However, these services have disadvantages for academic research because their rules are not transparent and therefore not intersubjectively verifiable. In addition, coding using geocoders can lead to many false positives (Takhteyev et al. 2012: 76). While a dictionary only finds exactly the terms contained in its lists, a geocoder relies on databases so comprehensive that almost any term can be included. This means non-geographic identifiers are often classified as locations.

Overall, there is no solution for the extraction and classification of location information in digital datasets that meets the criteria of transparency, reliability, and validity equally well. Researchers must consider the advantages and disadvantages of different methods in the context of the specific research question and data material.

2 Good coverage for major cities around the world is offered by, for example, the *iDAI* gazetteer of the German Archaeological Institute (https://gazetteer.dainst.org/ [last accessed: October 30, 2019]). In parallel, computer-assisted text analysis has developed tools aggregated under the term *named entity recognition* that can extract objects, including places, from text. One example is *SpaCy* (https://spacy.io/api/annotation#named-entities [last accessed: October 30, 2019]).

3 Geocoders and gazetteers: Two use cases and solutions

To illustrate the use, advantages, and disadvantages of the methods described, we present two examples from our research about the spatial location of *Twitter* networks. In one example, we demonstrate how to work with an automatic geocoder. In the second example, we identify location data using a dictionary created specifically for the research purpose. Both analyses were performed using the statistical programming environment *R* (R Core Development Team 2019).

3.1 Geocoder for systematizing location information

In our study of the spatial dimension of social networks, we investigated the digital interaction network around the city of Berlin. We wanted to know where in the world users who communicate with Berliners on *Twitter* are located. To map the geographical context, we had to standardize the inconsistent information from *Twitter* user profiles and assign coordinates to it. The dataset contained tweets by users who identified Berlin as their location and posted tweets in a two-week period in the summer of 2018. It also included all users with whom they interacted in their messages (through mentions, replies, retweets, and quotes). Overall, we had to classify the profiles of roughly 231,000 users.

Two reasons motivated us to use an algorithmically optimized geocoder: (1) The references to places were almost unlimited in their geographical and linguistic diversity. Such complexity is almost impossible to represent with existing gazetteers or even specifically constructed dictionaries. (2) As we had decided to analyze the location field, we could generally count on encountering usable place references. After comparing multiple service providers, we chose the German provider *OpenCage*, which is primarily based on *Open Street Map* data.[3] *OpenCage* offers monthly flat rate pricing models. After registering on the website, we used the API to submit queries using an authentication code. To do this, we used the *R* package *opencage* (Salmon 2018). One advantage was that the task required only minimal programming knowledge.

The *Twitter* dataset in our analysis stored the profile information for roughly 230,000 users in a number of variables (e.g., user name, description, number of followers, etc.). We extracted the location variable relevant for our study and geocoded them using *OpenCage* via the API. The result of each query was numerous structured annotations for each location, such as a formatted address, latitude and longitude, time zone, and individual variables representing geographical units at the level of neighborhoods to continents.

Data standardized in this way made it possible to create, for example, maps. We demonstrate this in Figure 2, which shows the most common locations of people connected to Berlin *Twitter* users at the level of cities (Fig. 2a) and countries (Fig. 2b). Further analysis, such as network analysis or linking locations to message content, is also possible.

3 https://opencagedata.com/ (last accessed: September 5, 2019).

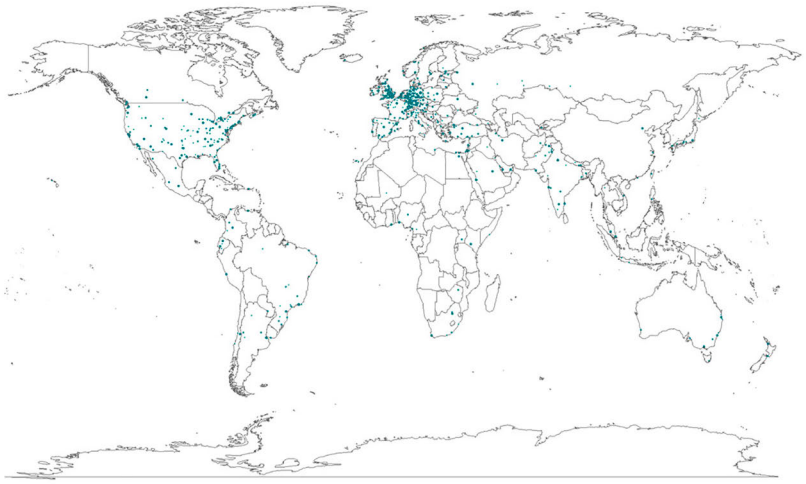

Fig. 2a: Locations of Twitter users connected to Berlin by city. |©Author's own diagram

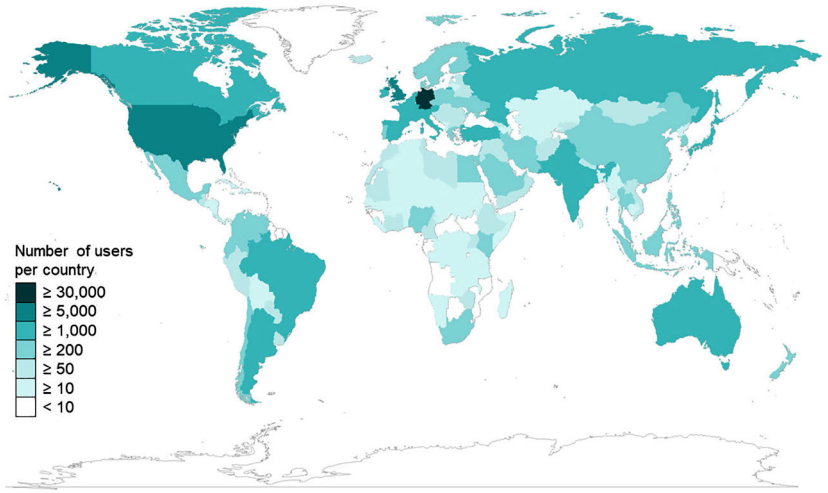

Fig. 2b: Locations of Twitter users connected to Berlin by country. | © Author's own diagram

Manual coding of a random sample of our dataset confirms that the chosen method produces valid results overall. Human coders judged the location coding identified by the geocoder to be correct in over 85 percent of cases. The most common type of error was false positives, in which the geocoder flagged a location but the human coder did not.

3.2 Coding with a specially constructed dictionary

In our second example, we aimed to extract references to places from running text. We used a dataset of around 50,000 German-language tweets concerning the Berlin housing market. What locations in the city were being discussed in this context?

Answering this question required a different approach than the first example. As that example showed, the main source of error in the geocoder was false detection of references to locations where there were none. If we had broken down the running text into individual words and coded them with the geocoder, we would have gotten a large number of false positives. The solution here was an analysis using a dictionary we prepared specifically for the project. It would be easy to expand classification to, for example, the extraction of streets or city neighborhoods. In the present case, however, we limited ourselves to references to Berlin's 96 urban districts. We performed two basic steps: first the preparation and then the application of the dictionary. We used the *R* package *quanteda* (Benoit et al. 2018), which contains a repertoire of functions for the automatic content analysis of text.

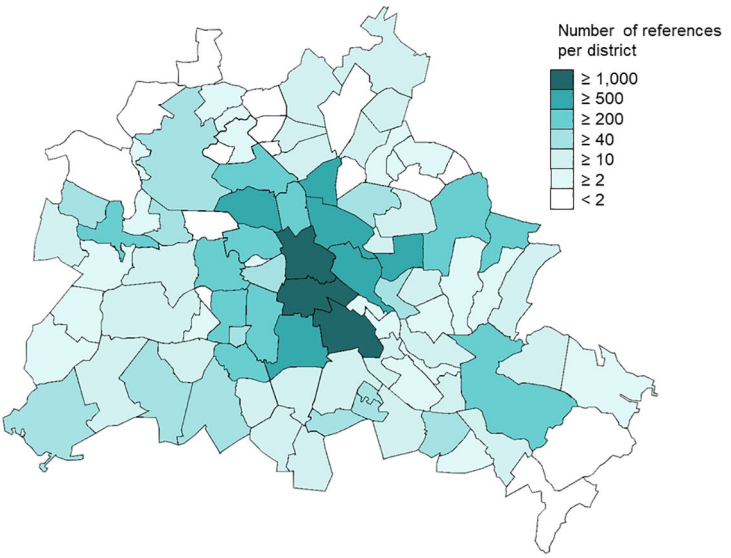

Fig. 3: Frequency of mentions of Berlin's urban districts in tweets concerning the housing market. | Map of Berlin's urban districts adapted from Angr (2007) for Wikimedia Commons.[4]

4 KarteBerlinDistricts.svg from Angr (2007) for Wikimedia Commons at https://commons.wikimedia.org/wiki/File:BerlinDistricts.svg (last accessed: April 29, 2020), licensed under CC BY-SA 3.0. Labels and colors have been removed from the original, the border of Borsigwalde district has been added, and districts have been colored to match the frequency distribution. The figures are likewise covered by the licensing conditions of CC BY-SA 3.0.

The most labor-intensive step in this method is preparing the dictionary. Researchers must specify what variations a place reference can have. A reference could be, for example, in uppercase or lowercase letters, with or without a hyphen, in the form of one or more words, or a hashtag. Abbreviations or inflections (e.g., "Pankow's residents") in which prefixes or suffixes are added to the place name can also make identification more difficult. The dictionary had to be constructed to reflect these variations.

The analysis itself was based on an existing function using only a few lines of code and took only seconds. The result was a *document feature matrix* (DFM), which records which urban districts are mentioned and how often for every tweet in our dataset. In our case, we found that 21.6 percent of all tweets referenced at least one district. This information can be displayed graphically, for example in the form of a simple frequency distribution (Fig. 3). Here, too, further analysis could consider, for example, the discourse relating to individual districts.

4 Research ethics

In this article, we have shown that digital data contains a multiplicity of location information, which opens up new perspectives for qualitative spatial research. The choice of the specific geocoding instrument depends on both the research question and the nature of the georeferences. While automatic geocoding has proven to be a valid method for classifying location information, dictionaries are better suited for the targeted analysis of individual georeferences in longer texts. The added value of digital geodata in the social sciences comes from triangulation and combination with traditional empirical methods. It opens up fresh perspectives in spatial research that are far from exhausted.

However, these opportunities for obtaining and analyzing large amounts of geoinformation also raise questions about research ethics. Although unimaginably large resources for scientific research are opening up as a result of this, location information is still observational data. The collection and analysis in aggregated form, creation of movement profiles of specific groups, and linking of the communication content to location profiles are generally done without the explicit consent of the users. Users consent to the collection of data in the general terms and conditions of the platforms and search engines. As long as they do not deactivate the location functions or location recognition of their device, these data resources can be misused for monitoring and *social scoring* (see Schütte/Klein 2020: 633). Calls for transparency to service providers whose business models are based on data with digital location information are therefore highly justified. Scientists who rely on access to the data should meet high standards of research ethics, methodological rigor, and compliance with data protection obligations. Only a transparent scientific process can justify using sensitive individual observational data in the service of social science research. This is true of all digital data, but it is especially true of digital location information because of the existing opportunities to link this information to social data.

References

Benoit, Kenneth/Watanabe, Kohei/Wang, Haiyan/Nulty, Paul/Obeng, Adam/Müller, Stefan/Matsuo, Akitaka (2018): Quanteda: An R Package for the Quantitative Analysis of Textual Data. In: *Journal of Open Source Software*, 3(309), pp. 774.

Cairncross, Frances (1997): *The Death of Distance: How the Communications Revolution will Change our Lives*. Boston, MA: Harvard Business School Press.

Duchastel, Jules/Laberge, Danielle (2019): Beyond the Quantitative and Qualitative Cleavage: Confluence of Research Operations in Discourse Analysis. In: Scholz, Ronny (Ed.): *Quantifying Approaches to Discourse for Social Scientists*. Cham: Palgrave Macmillan, pp. 23–47.

Goodchild, Michael F./Hill, Linda L. (2008): Introduction to Digital Gazetteer Research. In: *International Journal of Geographical Information Science*, 22(10), pp. 1039–1044.

Gore, Ross J./Diallo, Saikou/Padilla, Jose (2015): You Are What You Tweet: Connecting the Geographic Variation in America's Obesity Rate to Twitter Content. In*: PLOS ONE*, 10(9). Online: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0133505 (accessed: 28. Mai 2020).

Hecht, Brent/Hong, Lichan/Suh, Bongwon/Chi, Ed H. (2011): Tweets from Justin Bieber's Heart: The Dynamics of the ›Location‹ Field in User Profiles. In: *Proceedings of CHI 2011*, Session: Twitter Systems, Vancouver, pp. 237–246.

Hoffmann, Matthias/Heft, Annett (2020): »Here, There and Everywhere«: Classifying Location Information in Social Media Data-Possibilities and Limitations. In: *Communication Methods and Measures*, 14(1), pp. 1–20.

Hollstein, Betina/Straus, Florian (Eds.) (2006): *Qualitative Netzwerkanalyse.* Wiesbaden: VS Verlag für Sozialwissenschaften.

Kinsella, Sheila/Murdoch, Vanessa/O'Hare, Neil (2011): »I'm Eating a Sandwich in Glasgow«: Modeling Locations with Tweets. In: *Proceedings of SMUC 2011*, Glasgow, pp. 61–68.

Malik, Momin M./Lamba, Hamank/Nakos, Constantine/Pfeffer, Jürgen (2015): Population Bias in Geotagged Tweets. In: *Standards and Practices in Large-Scale Social Media Research. Papers from the 2015 ICWSM Workshop*. Palo Alto, CA: The AAAI Press, pp. 18–27.

McGee, Jeffrey/Caverlee, James/Cheng, Zhiyuan (2013): Location Prediction in Social Media Based on Tie Strength. In: *CIKM' 13: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, San Francisco, CA, pp. 459–468.

Rout, Dominic/Bontcheva, Kalina/Preoţiuc-Pietro, Daniel/Cohn, Trevor (2013): Where's @wally? A Classification Approach to Geolocating Users Based on their Social Ties. In: *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, Paris, pp. 11–20.

R Core Development Team (2019): *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Russom, Philip (2011): Big Data Analytics. In: *TDWI Best Practices Report*, 19(4), pp. 1–34.

Sadilek, Adam/Kautz, Henry/Bigham, Jeffrey P. (2012): Finding Your Friends and Following Them to Where You Are. In: *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, Seattle, WA, pp. 723–732.

Salmon, Maëlle (2018): *Opencage: Interface to the OpenCage API*. R Package Version 0.1.4. Online: https://docs.ropensci.org/opencage/ (accessed: 28. Mai 2020).

Scharkow, Michael (2013): Automatische Inhaltsanalyse. In: Möhring, Wiebke/Schlütz, Daniela (Eds.): *Handbuch standardisierte Erhebungsverfahren in der Kommunikationswissenschaft*. Wiesbaden: Springer VS, pp. 289–306.

Schütte, Hendrik/Klein, Maximilian (2020): Social Credit Ratings in der Praxis – dargestellt am Beispiel des Wareneinkaufsfinanzierers. In: Everling, Oliver (Ed.): Social Credit Rating. Wiesbaden: Springer Gabler, pp. 627–639.

Takhteyev, Yuri/Gruzd, Anatoliy/Wellman, Barry (2012): Geography of Twitter Networks. In: *Social Networks*, 34(1), pp. 73–81.

Wilken, Rowan (2014): Twitter and Geographical Location. In: Weller, Katrin/Bruns, Axel/Burgess, Jean/Mahrt, Merja/Puschmann, Cornelius (Eds.): *Twitter and Society*. New York, NY: Peter Lang, pp. 155–168.