Evaluating the Use of Gazetteer Data for Locating Former Concentration Camp Subcamps

Francis Harvey[1] and Ihor Doroshenko[1]

[1]Leibniz-Institut für Länderkunde, Germany

Abstract

How can researchers use available gazetteer data to determine and assess the locations of former Nazi concentration camps and their subcamps? Often these places are known only by the names of the towns, villages, or areas of cities where the subcamps were (possibly) located. In this paper, we consider the challenges of using digital gazetteer data to help identify the actual locations of the camps. We assess the accuracy of the locations we can establish, comparing these to the locations of the subcamps available from the Buchenwald Memorial Foundation. Researchers can exploit gazetteer location data using GIS or online mapping tools. Using publicly available GIS methods and various geo-statistical techniques, we compare place names associated with concentration camp subcamps with historical and contemporary digital gazetteer information. The results suggest that gazetteer data can help to map larger areas, but are often ill-suited for precise identification of sites of remembrance. This lack of certainty points to the need for further information and to the complexities of reliably establishing locations using geographic information from geomedia.

Keywords:

Geocoding, remembrance sites, Holocaust, positional accuracy

1   Finding sites of remembrance

To locate sites of remembrance known only by place name, perhaps only the name of a town, researchers and interested people can turn to various technologies using geomedia, such as GIS, GPS, and offline- and online-mapping applications. The emphasis on historical memory is greatest in historical research and in relation to mourning connected to specific places. People remember places even if they cannot be located accurately. In this contribution, our focus is on the technological aspects of connecting places of memory (concentration camps' subcamps) with the actual locations of the subcamps. We consider the geographical accuracy that various types of digital gazetteers and place names can achieve for locating subcamps reliably. The information in such sources is frequently collected as volunteered geographic information and as such may not always be 100% accurate. This type of geospatial data is exemplified by GPS-data collections and institutional gazetteer data. We use digital data for the locations of subcamps (Aussenlager) of the Buchenwald Concentration Camp which were

made available by the Buchenwald/Mittelbau-Dora Memorials Foundation. The geographical and thematic focus of our article speaks to the possibilities and challenges of using geomedia for place-based remembrance. In particular, we are concerned with limitations in accurately finding specific places from sparse historical information and the challenges this poses. We show that using information from multiple sources provides reliable ways to understand the discrepancies between data on subcamp locations and gazetteer data and possibly to find ways to improve accuracy in pinpointing sites.

The United States Holocaust Memorial Museum Encyclopaedia 2009 identified 20,000 camps, subcamps and ghettos. That was updated in 2013 to 40,000, and in 2016 to 42,500 (Megargee, 2009–2018). Still, locating specific sites of concentration camp subcamps can be a challenge. Roz Vara and Charlie Hoffman located 158 camps but did not give their geographical coordinates (Knowles, Cole and Giordano, 2014). Benjamin Perry Blackshear improved the geographical accuracy of other camp locations. He triangulated latitude and longitude by comparing the geographical description of locations from an encyclopaedia with the locational data from digital gazetteers, satellite imagery and online maps (using Google Maps' Latitude Longitude popup information), and getlatlon.com (Knowles, Cole and Giordano, 2014).

To illustrate the possibilities and challenges of identifying the discrepancies referred to above, we compare subcamp locations with some currently available digital gazetteer information. We have not been exhaustive in considering all possible information sources, for example analog records from researchers and remembrance organizations (e.g. the European Historical Research Institute (EHRI)).

The remainder of our contribution is divided into three sections. The next section covers the possibilities and challenges of locating sites of remembrance as a form of geocoding that uses specific historical data sources. The following section presents an analysis of positional accuracy which we conducted by contrasting the locations of subcamps of the Buchenwald concentration camp using two digital gazetteers, GeoNames and Google Earth. The concluding section provides a summary and vital points for using historical data and digital gazetteer data to find sites of remembrance.

Throughout, we use the generic term 'populated places', as found in censuses and gazetteers, to refer to sites of persistent human inhabitation. These include larger communities such as cities, towns and villages.

2   Potentials and Challenges of locating sites of remembrance

This section covers the possibilities and challenges of locating sites of remembrance as a form of geocoding that uses specific historical data sources. Generic issues of geocoding in locating places or sites from names or from address information are further complicated by the complex, often imprecise, and possibly also obscured geographical information. The essence of the geocoding process is understanding how information about place can be most accurately connected to specific locations for the intended purpose. It involves complex space/time challenges with similar methodological challenges which, when diligently pursued, open up significant possibilities for deepening historical and geographic understanding. Our focus here

is on connecting places to locations, which can be mapped using widely available applications. These can be simple online mapping of locations as a series of steps, or more complex GIS-based geovisualizations that involve cartographic and geographic information transformations. (How to use different mapping packages is outside the scope of this article.)
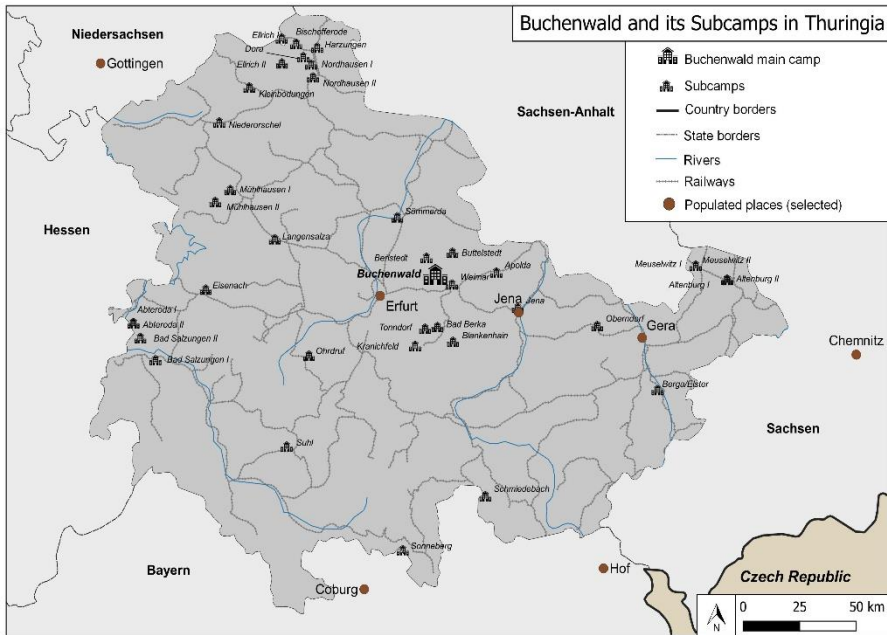
3   Potentials and Challenges of Geocoding

Geocoding, the process of connecting a place with geospatial information (coordinates), helped GIS become a success in the 1990s: it was now possible to take historical place descriptions, connect them to locations, and analyse geographical relationships. GIS became an essential part of many digital or spatial humanities research projects (Bodenhamer, Corrigan and Harris, 2010), and was used for any amount of data, large or small. The locations where individual letters were written (The Valley Project, no date) or thousands of records from ship logs (Farrington, Lubker, Radok and Woodruff, 1998) could now be visualized and digitized, and used, for example, for the analysis of wartime movements or in research into global changes. Successful geocoding was also the gateway to carrying out more complex spatial analysis (Chrisman, 1997) and using spatial statistics (O'Sullivan and Unwin, 2003). In contrast to digitization, which was the primary way to create digital data for GIS, geocoding has continued to be an effective method for transforming place information into locational data.

It is worth reminding readers of some of the subtle but impactful geodetic complexities that can bedevil geocoding, for instance, datums (i.e. coordinates). As the earth moves through the solar system, tectonic forces, both subtle and dramatic, lead to changes in the shape of the earth, as do the movements of magnetic fields and poles. These changes mean that locations become 'displaced', and thus specific datums need to be updated. People mapping at different times may knowingly or unknowingly use different datums for the same location. It is thus possible that the same site will appear in different places in a mapping geovisualization. To this challenge are added those of map projections, which map a two-dimension coordinate system on to the roughly spherical shape of the earth. Even avoiding projections by working with degrees (for example using GPS degrees, minutes, seconds (DMS) constraints), or by limiting the number of decimal places recorded with the data, may significantly impact accuracy or precision (Chapman, 2005).

For many people working with geographic information, these technical issues often appear obscure, presented in very technical language. Wieczorek et al. (2004), however, make a significant contribution to discussion of issues that influence positional accuracy when transforming place names into mappable location data, pointing to six sources of uncertainty that impact geocoding:

1) extent of the locality
2) unknown datum
3) imprecision in distance measurements
4) imprecision in direction measurements
5) imprecision in coordinate measurements
6) map scale.

Using these elements, it is possible to break down the complexity of location information. While some of these sources of uncertainty are unlikely to be relevant for some uses of data for locating sites of remembrance, all are likely to be inherent in the description of places. They serve to remind us that detailed locational data is not by itself accurate, but understanding the limitations can help us assess the data accuracy for a specific (intended) purpose (the fitness for use paradigm) (Chrisman, 1982).

Figure 1: Location of the subcamps of Buchenwald and Dora(-Mittelbau) concentration camps in the present state (Bundesland) of Thuringia

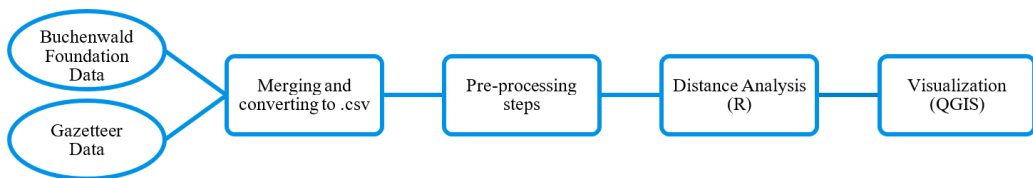# 4   Using digital gazetteers to locate former concentration camp subcamps

It is regrettably not possible in the scope of this study to assess all six aspects of data accuracy as defined by Wieczorek et al. (2004). Instead, we address a more straightforward question: what differences are there between the locations of places that historically or currently bear names associated with particular subcamps, and the Memorial Foundation's locations of the those same Buchenwald concentration camp subcamps (e.g., subcamps Altenburg I and Altenburg town)?

5   Background

One website showing locations of the Dora(-Mittelbau) subcamps, http://www.aussenlager.dora.de/l/de/home/map, and information in five languages offers much specific information. Spatial and temporal issues described on the website impact how accurately we can determine the location of the subcamps. As the notes for individual sites explain, the exact location of a subcamp may be unknown; what we have by way of evidence is only the name of the populated place of the same name, and when the name was first used. Subcamps may also have been used only periodically, or only for a limited period. Because of missing documents, it may not be possible to ascertain the precise facility or the exact location of a subcamp. We assume for this analysis that the Buchenwald Foundation data is accurate and verified.

6   Analysis

In the following, the analysis steps are described (see Figure 2). The 37 Buchenwald subcamps, all of which were in what is now the German state (Bundesland) of Thuringia, were identified for us as point locations, as shown in the Appendix (Table 1). We created a table, with coordinates from the PDFs provided to us, showing the location and other information for each subcamp. After collecting the gazetteer data, we extracted locational data for the place names of subcamps from four data sources: GeoNames (one of the most widely-used digital gazetteers); GoogleMaps; and two historical sources (the Amtliches Ortsverzeichnis (1944) and the Post-Taschen-Atlas (1917)) which we digitized. A GIS neighbourhood analysis was used to detect the place closest to the location of each subcamp, as defined in the data from Buchenwald concentration camp. Of course, the localization was possible only when spatial coordinates were available, which was not the case for historical gazetteer data. It was very important to consider not only modern data but also historical data, in order to avoid misinterpretation of the results due to the temporal gap between the Holocaust-related and modern spatial data, and of the possible inaccuracies in the datums, as described earlier.

Figure 2: Workflow of the analysis

A text-string distance function was used to detect the closest possible corresponding place names in the two historical sources. Textual distance functions map two strings and return a value of difference, which is small if words (or in our case toponyms) are similar, and larger the more they differ. An important part of distance functions is considering so-called 'edit
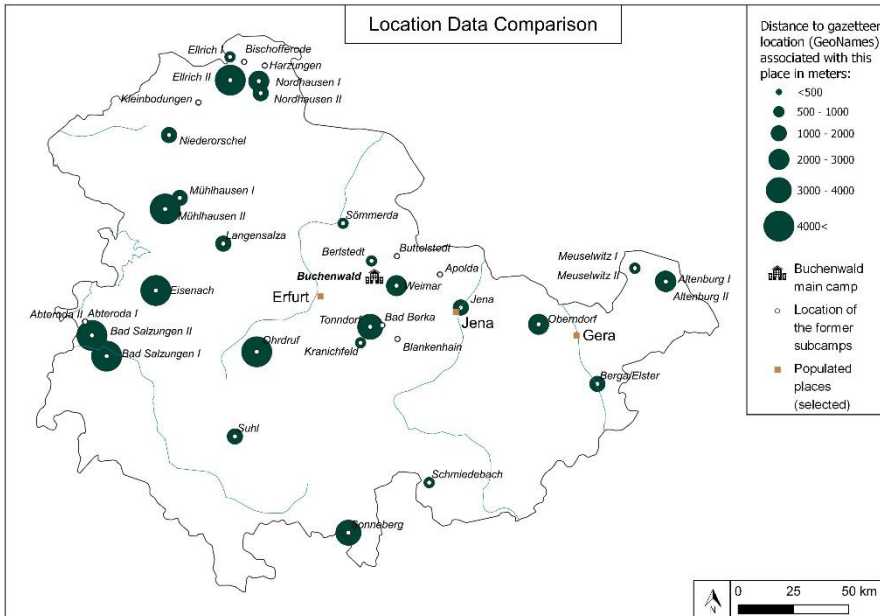
distances', which measure the hypothetical costs of converting one word to another (Cohen et al., 2003). Because the goal is to detect difference between words, the Levenshtein edit distance was selected as the best-documented and most widely used for linguistic research (Kessler, 1995; Angelis et al., 2017). The algorithm finds the 'cheapest' way of converting one word into another, checking all possible combinations. The words (or names) are matched against each other and the minimum cost is calculated. The cost consists of three operations (deletions/insertions/substitutions of individual letters, and a value of 1 is given to each operation performed. After that, the values are summed, resulting in the Levenshtein distance between two strings (see Figure 3).

String 1		k	r	y	s	z	b	o	r	k	
String 2	c	h	r	i	s	t	b	u	r	g	
Costs of conversion	1(i)	1(s)		1(s)		1(s)		1(s)		1(s)	= 6

Figure 3: Example of the Levenshtein distance calculation (s – substitution; i – insertion)

After the textual pre-processing, a spatial multi-source place-name database was created, consisting of the subcamp names, names of reference populated places, and spatial coordinates. This data was used for the analysis of locations of subcamps and corresponding populated places, and for the analysis of the exact textual and spatial discrepancies between the names from the data sources (discussed below). First, the text string-distance was measured, using the Levenshtein distance function, and stored in the form of a matrix, with the names of subcamps on the y-axis and names of sources on the x-axis. The minimum distance was 0, if there was no difference between names (for example, Abteroda and Abteroda); the maximum difference was 17 (between Oberndorf and Bad Klosterlausnitz). Some pre-processing was done to reduce the detection of unnecessary string elements, such as additional geographical information in brackets (for example, 'Provinz Sachsen' was deleted as it was not a part of the name itself). However, numbers (in names of subcamps like Altenburg I or Altenburg II) were preserved, because they serve to differentiate subcamps. Finally, the spaces between elements in composite names (e.g. in Bad Salzungen) were deleted.

Secondly, the spatial distance between subcamps and the locations of similarly- or identically-named populated places was measured using the 'geosphere' package. The package is used for spherical trigonometry for geographic applications. It computes distances and the latitude and longitude of locations. The package's distHaversine function was used to retrieve the shortest distance (the 'great-circle-distance' or 'as the crow flies') in metres between two points. This method assumes a spherical earth, ignoring ellipsoidal effects (Hijmans, 2019). Because of the lack of coordinates in the historical data, the possibility of working directly with the database, and the effort required to support open-access data, only toponyms found in GeoNames were covered in this part of the analysis.

Figure 1: Distances between the Buchenwald concentration camps and subcamps, and corresponding places from the GeoNames gazetteer for the present state (Bundesland) of Thuringia

Figure 4 shows the spatial differences in metres between the subcamp locations (data from the Buchenwald memorial site) and GeoNames. Five of the 37 subcamps' actual locations are more than 4 km from the corresponding GeoNames locations. Discrepancies for the majority of the other subcamp locations are between 500 and 2,000 metres.

The larger discrepancy still allows GeoNames-based locations to be used for making maps of the region. However, with such a large discrepancy, it is not possible to zoom in on the maps. For online browsing, zooming in to a specific town may lead to erroneous conclusions regarding the site of the subcamp. The five sites with a difference of more than 4,000 metres from GeoNames would be unsuitable even for many regional maps; clear information regarding the inaccuracy of their locations would assist map readers in assessing the suitability of particular maps. In one case, the distance is over 16 km. Research into the sources (following Wieczorek et al.'s criteria) of the locational discrepancy would be helpful for understanding it.

Using string-distances, it was possible to detect text-based discrepancies, because all sources used provide names of places. Figure 5 shows the differences between the names from the gazetteer in question and those provided by the memorial site. It is notable that the most significant differences are between place names from modern sources (Google-Maps and GeoNames) and the memorial sites' data. Furthermore, a significant difference was also detected for the 'Post-Taschen-Atlas', which was published almost 20 years before the first of the subcamps researched here was built. In contrast, the smallest distance was detected for 'Amtliches Ortsverzeichnis', which was published in 1944 and is thus the closest source, temporally speaking, to the time when the places became subcamps and received their names.

One could observe here a correspondence between the string distance and the temporal remoteness of a source from which a place name was derived. Subcamps received official names, which often but not always (as we can see in current work) corresponded to some extent to names of neighbouring places, which were in a particular administrative unit. It could be, then, that over time administrative units changed, which then affected the association of place with subcamp location. So using data from historical gazetteers could produce erroneous results or be a source of uncertainty. This temporal relationship requires further research, using additional sources.

The present research, however, is only our first attempt to perform this analysis. The limitations we faced are also relevant, as they may reflect more general issues for research into subcamps. The actual data (place names) were derived from sources which do not necessarily reflect the official historical names, but rather are place names used by gazetteer makers at the time. Consequently, the results show textual and spatial relationships between names of Buchenwald subcamps and places from non-official sources. To understand how far or close the official place names were in relation to the subcamp names, one has to work with primary sources such as acts and documents rather than with gazetteer data. Unfortunately, the gazetteer data, especially the historical data, often reflect subjective perspectives and interpretations.

Second, the spatial resolution of the data in the gazetteers is an important issue. Historical sources, unlike modern ones, cannot provide detailed information for locations, as they do not have, e.g., a zooming function. This is why information from many sources is only approximate, and a subcamp, for example, might be associated with a larger or more important populated place, and not necessarily with the nearest one.
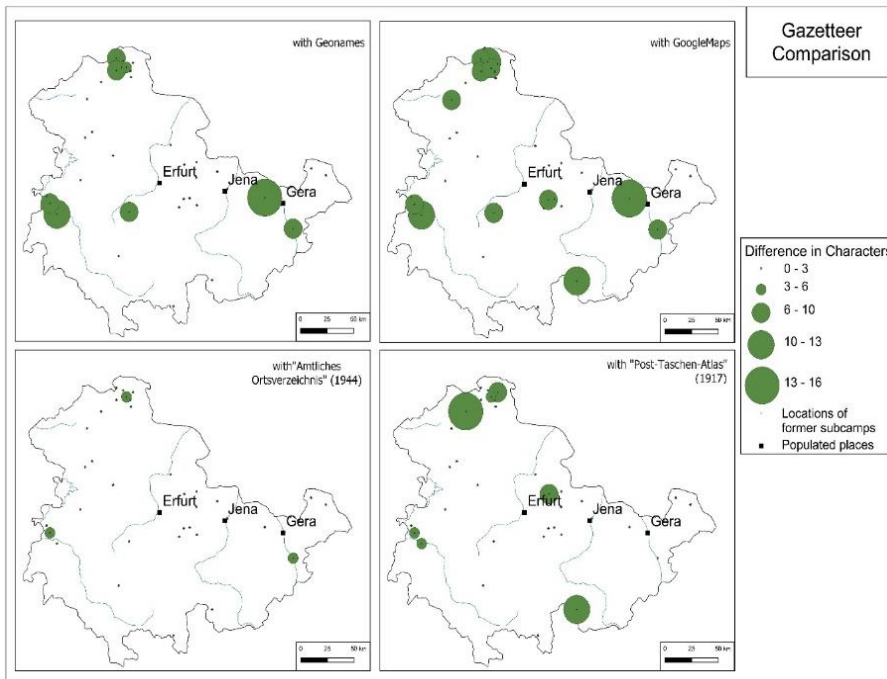
Third, Thuringia as a region did not exist at all in 1917, the year when the first of our sources was published, although the spatial dimension of the research does appear to coincide with the extent of Thuringia's modern-day borders. To avoid a certain amount of data processing, one has to consider either (a) using a different geographic area and focusing exclusively on the entire Buchenwald camp system, or (b) using only those sources which are applicable for present-day Thuringia. But even in this case the borders can be open to doubt, because they have not been stable over the period from when the subcamp was built to the present day.

Fourth, the methodology to detect the places that correspond most closely in textual and spatial senses should be refined with the addition not only of buffer and intersection tools, but also of statistical approaches such as Nearest Neighbour Analysis to provide statistically more reliable results. Finally, as already mentioned, additional sources are needed, which could reveal trends and relationships which cannot be detected using our 4 sources, only 2 of which have spatial coordinates.

Fifth, the techniques used to detect textual difference do not always reflect how place names may (or may not) differ from each other. As we know, language is not just written, but also pronounced or spoken. The perspective taken in the current work tends to understand language only as something written, and distances measured are based only on different spellings and not on a comparison of their phonetic forms. With the help of the Soundex algorithm, for example, one can detect the differences between words as they are *pronounced*, like the Levenshtein algorithm does with regard to spelling.

As both parts of our analysis show, associations of subcamp locations with populated places can lead in some cases to erroneous conclusions. One might think that subcamps were located close to the places which bear the same names as the camps, but this is not true in all cases. This analysis may serve as the starting point for deeper historiographical research exploring how people see the sites of former concentration camps, and just how they can be located.

Figure 5: Differences in number of characters for the names of subcamps considered in this paper.

7   Conclusion

What this analysis shows is that the locational accuracy for subcamps is in many cases uncertain to a significant degree. The digital and historical gazetteer data we utilized is frequently not sufficiently reliable to connect a place of remembrance with the actual location of events. For most uses, the accuracy is still generally sufficient for larger areas (maps of, e.g., regions). However, for research using locational data available) in geomedia, adequate information on metadata and the presentation of data should be offered. Possibly, even, the display of data for smaller areas should be limited (in order to avoid overloading maps visually). This latter point is relevant also to assessing the accuracy of geocoding for volunteered data collections (Goodchild, 2007) and newer forms of data integration that digital geoinformation makes possible, e.g. collaborative geocoding.

To conclude, we can highlight several points from the analysis that also suggest some directions for further research. First, the discrepancies in distances between gazetteer locations

and data from the Buchenwald memorial suggest that while for the most part the gazetteer-based locations would be sufficient for mapping sites at regional or national scales, the differences limit the use of the gazetteer data for detailed mapping (e.g., to show sites accurately on a local map) (see Figure 5). Second, knowing the purpose of the mapping is essential to assess the desirable degree of locational accuracy. Using Wieczorek et al.'s list of factors for geocoding would provide valuable information to guide the assessment of locational accuracy. Third, also following Wieczorek et al. (2004), using and presenting subcamp locational data should come with clear information about its accuracy and documentation on how the data was collected.

As pointed out in the introduction, locating places of remembrance is complicated. The place remembered is difficult to connect to the location of specific events. This is a significant research area for historians and geographers, and more recently for those working in digital humanities. Looking at degrees of linguistic similarity among toponyms for Buchenwald subcamps in Thuringia suggests that transliteration and other linguistic differences should be taken into account **(**Bodenhamer, Harris and Corrigan, 2013). In this analysis, we compared simply the number of characters used in different data sources. While this gave some interesting results, they need further interpretation, and a method is required that can be used to identify differences between toponyms for the same places in extensive data collections (Figure 4).

While this article makes only a small contribution to the challenges of locating and mapping sites of remembrance, hopefully its consideration of locational data points for concentration camp subcamps gives an indication of essential issues and geocoding guidelines that can help historical researchers using gazetteer data better to associate places with the locations of events. Explaining the issues that can arise from using information in gazetteers will hopefully aid future research in locating additional places of remembrance, and in locating them more accurately.

Acknowledgements

References

Angelis, G. de, Jessner, U., & Kresic, M. (Eds.) (2017). Crosslinguistic influence and crosslinguistic interaction in multilingual language learning (Paperback edition). London: Bloomsbury Academic.

Bodenhamer, D., Corrigan, J., & Harris, T. M. (2010). The spatial humanities. Bloomington: Indiana University Press.

Bodenhamer, D. J., Harris, T. M., & Corrigan, J. (2013). Deep mapping and the spatial humanities. *International Journal of Humanities and Arts Computing*, *7*(1-2), 170-175. Retrieved from https://diginole.lib.fsu.edu/islandora/object/fsu:209941/datastream/PDF/view

Chapman, A. D. (2005). *Principles of data quality*. Copenhagen: Global Biodiversity Information Facility.

Chrisman, N. R. (1982). A theory of cartographic error and its measurement in digital data bases. Proceedings from Auto-Carto 5.

Chrisman, N. R. (1997). *Exploring Geographic Information Systems* (First ed.). New York: John Wiley & Sons.

Cohen, W., Ravikumar, P., & Fienberg, S. (2003). A Comparison of String Distance Metrics for Name-Matching Tasks. *American Association for Artificial Intelligence*.

Farrington, A., Lubker, S., Radok, U., & Woodruff, S. (1998). South Atlantic winds and weather during and following the Little Ice Age, A pilot study of English East India Company (EEIC) ship logs. *Meteorology and Atmospheric Physics*, *67*, 253-257. doi:10.1007/BF01277515

Goodchild, M. F. (2007). Citizens as voluntary sensors: Spatial data infrastructure in the world of Web2.0. *International Journal of Spatial Data Infrastructures Research*, *2*, 24-32.

Hijmans, J. R. (2019). *Package 'geosphere'*. Retrieved from https://cran.r-project.org/web/packages/geosphere/geosphere.pdf

Kessler, B. (1995). Computational dialectology in Irish Gaelic. In S. P. Abney & E. W. Hinrichs (Eds.), Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics – (p. 60). Morristown, NJ, USA: Association for Computational Linguistics. https://doi.org/10.3115/976973.976983

Knowles, A. K., Cole, T. and Giordano, A. (Eds.) (2014). *Geographies of the Holocaust*. Bloomington and Indianapolis: Indiana University Press.

Megargee, G. P. (2009-2018). *The United States Holocaust Memorial Museum Encyclopedia of Camps and Ghettos, 1933—1945*. Volumes. I - III. Bloomington and Washington, D.C.: Indiana University Press and in association with the United States Holocaust Memorial Museum. Retrieved from https://www.ushmm.org/research/publications/encyclopedia-camps-ghettos

O'Sullivan, D., & Unwin, D. J. (2003). *Geographic Information Analysis*. New York: John Wiley and Sons.

Pfuhl, T. (1917). Post-Taschen-Atlas von Deutschland: nebst Ortsverzeichnis. Frankfurt (Oder): Selbstverlag von Th. Pfuhl

The Valley Project. (no date). Valley of the Shadow. Retrieved 25Feb, 2020 from http://valley.lib.virginia.edu/VoS/lettersp2.html

Wieczorek, J., Guo, Q., & Hijmans, R. (2004). The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science*, *18*(8), 745-767.