SpaceWars: A Web Interface for Exploring the Spatio-temporal Dimensions of WWI Newspaper Reporting

Nicolas Gutehrlé[1], Oleg Harlamov[2], Farimah Karimi[3], Haoyu Wei[4],
Axel Jean-Caurant[5] and Lidia Pivovarova[4]

[1]*University of Franche-Comté, France*
[2]*Friedrich–Alexander University Erlangen–Nürnberg, Germany*
[3]*University of Cologne, Germany*
[4]*University of Helsinki, Finland*
[5]*University of La Rochelle, France*

Abstract

In this paper, we describe an interactive map that places automatically extracted location names on a map and allows historians to study spatial imaginaries across time and newspapers. Our contribution is two-fold: first, we present a working instrument for historical studies that is freely available on the web; second, we describe a data analysis pipeline, which can also be applied to other data and material. Challenges we address in this paper range from handling textual noise introduced by Optical Character Recognition (OCR) and Named Entity Recognition (NER) applied to historical documents, georeferencing textual place mentions, and issues connected with web design and the ultimate user experience.

Keywords

Named Entities, visualization, georeferencing, newspapers, WWI

1. Introduction

The number and size of digitized text collections is continually growing leading to a rise of interest in macroanalysis or distant reading methods[1] We offer a tool for historians that visualizes the locations mentioned in newspapers on a map and provides various functions to analyze spatio-temporal information, thus simplifying studies of spatial imaginaries.

Spatial imaginaries are ”stories and ways of talking about places and spaces that transcend language as embodied performances by people in the material world” [2]. Spatial imaginaries are shared by large groups of people, or a society as a whole [3]. Media play a vital role in reflecting public discourses, including spatial imaginaries. This could be seen in ways they are referring to different locations and, even more importantly, in the amount of attention they pay to various places. Huge events, such as wars, transform spatial imaginaries in many ways. This

HistoInformatics 2021 – 6th International Workshop on Computational History, September 30, 2021, online
nicolas.gutehrle@univ-fcomte.fr (N. Gutehrlé); oleg.harlamov@fau.de (O. Harlamov); karimi.farimah@gmail.com (F. Karimi); hayou.wei@helsinki.fi (H. Wei); axel.jean-caurant@univ-lr.fr (A. Jean-Caurant); lidia.pivovarova@helsinki.fi (L. Pivovarova)

transformation is reflected in and accelerated by the media and thus can be observed in large collections of historical newspapers.

Even though automatic analysis of massive text collections can help reveal hidden patterns in the data and lead to new research questions, it is not sufficient to study the construction of spatial imaginaries. As stated by [4], it is necessary to go back to the source document and study these location mentions in their context combining distant and close readings of the documents. Thus, in addition to maps that provide a summary view of the data, our tool provides concordances for locations and links locations to the articles mentioning them, allowing historians to alternate massive automatic analysis with close reading.

The newspapers annotated with automatically extracted locations are provided by the NewsEye project [5], aimed at providing analytical tools for historical analysis of the past media. We limit our tool with the period of the First World War. NewsEye collection contains newspapers from Austria, France and Finland, which belonged to different parties during WWI and got their news from different channels.

In most of the texts the locations have been extracted and tagged. However, macroanalytical tools within NewsEye are not centered around locations. Visualizing spatial entities from a text collection in a geographical information system is a type of digital media transformation that puts the focus on the geospatial aspects of the data. Our web app thus offers a feature that was missing from the NewsEye platform and that enables users to explore the perception and influence of space and place in relation to other dimensions of the Great War[1].

We describe all steps performed to develop a web tool. Our code is publicly available[2] thus our approach can be reused for other historical use-cases. The remainder of the paper is organized as follows: Section 2 introduces some theoretical studies in the field of Geo-Humanities and related mapping projects. In Section 3 the underlying dataset is described, Section 4 outlines the data preprocessing and in Section 5 the web interface is presented.

2. Related Work

There have been extensive theoretical studies in the field of spatial digital humanities and on the methods of geographical text analysis and digital mapping [6, 1, 7]. [1] focuses on the conceptual modelling and transformation of textual data to the medium of maps, while [7] offer a collection of essays on the influence of the use of geospatial analysis in the discipline of literary studies and showcases a number of literary mapping projects. [8] propose a methodology to extract ”spatial-temporal profiles” which list normalised temporal and spatial expressions that co-occur in documents. Such profiles can then be used to visualise on a map the succession of events as depicted in the document in chronological order. The ELEVATE-live web interface (https://elevate.greyc.fr/) by [9] shows the ”virality” of news article across countries through map visualization. This interface relies on previous works [10], which demonstrates that the inherent semantics of documents can be explored thanks to information at the entity-level.

Next to projects revolving around literary and fictional texts (e.g., *A Literary Atlas of Europe* (http://www.literaturatlas.eu/en/index.html), *A Map of Paradise Lost*(https://olvidalo.github.io/paradise-

[1] The web interface is available at http://spacewars.newseye.eu/

[2] https://github.com/dhh21/SpaceWars

lost/), there are also various examples of projects based on non-literary historical data. The *al-Ṯurayyā* Project(https://althurayya.github.io/) is a gazetteer and a geospatial model of the early Islamic world visualizing over 2,000 toponyms and route sections from Georgette Cornu's *Atlas du monde arabo-islamique à l'époque classique: IXe-Xe siècles* (Leiden: Brill, 1983) on an interactive map with additional path finding and search features. A similar but much larger project is *ORBIS: The Stanford Geospatial Network Model of the Roman World* (https://orbis.stanford.edu/), which "reconstructs the time cost and financial expense associated with a wide range of different types of travel in antiquity" [11].

Running Reality(https://www.runningreality.org/) is an expansive research project that aims to model the evolution of human civilization. The spatio-temporal system is based on a complex world history model and can render any day of any year as a map that allows users to interact with and gain information about the displayed spatial entities and objects. *Trading Consequences*(http://trcons.edina.ac.uk/vis/tradConVis/) uses text-mining software to explore historical documents related to international commodity trading in the British Empire during the 19th century, and its impact on the economy and environment. The web map utilizes a combination of text-based and graphic visualization techniques like maps, diagrams and word clouds to present the information extracted from the documents.

[4] applies distant reading methods to study how the *Houston Daily Post* between 1984 and 1901 shaped what he terms an *imagined geography* by privileging places over others. Blevins approximates this privilege ranking of place mentions via the corresponding occurrence frequencies in this newspaper collection(http://spatialhistory.stanford.edu/viewoftheworld). These maps revealed that local places such as Houston, Dallas or Austin are much more frequent in the *Houston Daily Post* than more thriving cities at the time such as New-York or Chicago. This suggests that, at a time where the United States underwent an important process of nationalization and standardization, the newspaper actively emphasized its regional space.

Similarly to [4], we developed a web interface to explore how newspapers published during WWI created spatial imaginaries. Through this app, the user can explore the whole dataset or focus on specific newspapers, languages or time periods. The interface consists of two parts which we describe below: the map module and the concordancer module. The main difference between our project and others is that our application directly connects occurrences of place names to their corresponding geospatial entities visualized on the map.

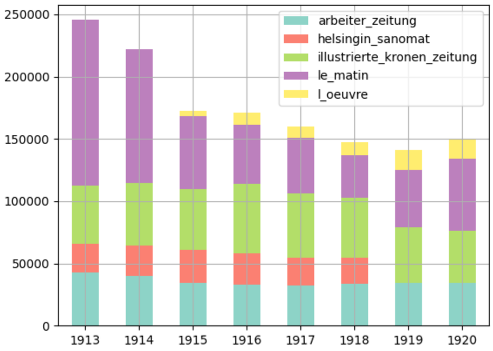
3. Dataset

3.1. NewsEye collection

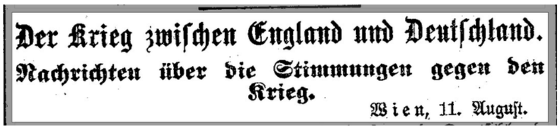
The underlying data[3] consist of automatically recognized and NER-tagged historical newspapers Whenever possible, extracted entities have been automatically linked to a Wikidata resource. The details of the method can be found in [12, 13], and other project publications.

We use a subset published between 1913 and 1920. It covers 5 different newspapers from 3 different European languages: *Arbeiter Zeitung*, *Illustrierte Kronenzeitung* (German), *Le Matin*,

[3] The data have been provided by courtesy of the *NewsEye* project: https://www.newseye.eu/; for details, see [5]
.



(a) Unique toponym counts per newspaper-year



Der Krieg zwiſchen England und Deutſchland.
Nachrichten über die Stimmungen gegen den Krieg.
Wien, 11. Auguſt.

(b) Excerpt from the *Neue Freie Presse*, 11.08.1914

L'Œuvre (French), *Helsingin Sanomat* (Finnish). The number of unique location names for each year is shown in Figure 1a. Note that this statistics reflects peculiarities of the dataset rather than historical processes directly: e.g. *L'Œuvre* for 1913 and 1914, and *Helsingin Sanomat* for 1919 and 1920 are missing even though the newspapers were published in these periods. We still need to explain the larger number of unique names in 1913 and 1914; one possible explanation might be a worse OCR quality that leads to producing more name variants.

The quality of the OCR and NER varies across newspaper issues and languages, which creates major challenges for assessing the accuracy of the character strings recognized and classified as place names. For example, in Figure 1b we show an excerpt from an Austrian newspaper article announcing a war between Britain and Germany. However, the NER system recognized only one location in this text, namely Vienna, which is the publication location rather than a location of the event. For NewsEye collection, the NER F-measure varies from 57.13% (French) to 36.39% (Finnish), the Named Entity Linking (NEL) F-measure varies from 48.4% to 30.0%. However, the numerical results were obtained using small subsets of the data, since there are no manually annotated historical datasets for these tasks. Such problems are common for historical datasets [14, 15] and explains why automatic processing is usually accompanied by close reading.

3.2. External reference

To compare location coverage in media with places where events took place in reality, we completed our dataset with data relating to battles during the WWI.

We are unaware of any professional database providing access to battle locations in a machine-readable form and thus rely on Wikidata to obtain this information. Wikidata contains well-structured battle informations and saves them as a semantic frameworks. We extracted battles from Wikidata using SPARQL with the *WWI* keyword as a query. For each battle, we extracted its name and coordinates, start and end dates, the war fronts it belongs to (e.g. Western), the country where it took place, and its duration in days. The ontology of war fronts on Wikidata is illustrated in Figure 2. The WWI ontology has a structured hierarchy, most of the battles are linked to a specific front, some are linked to WWI directly, and some of the battles are linked to

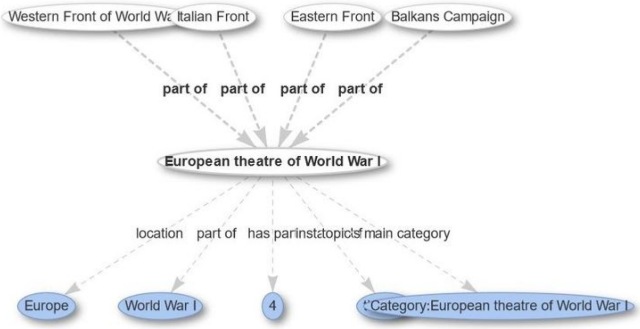

Western Front of World Wa
Italian Front
Eastern Front
Balkans Campaign
part of
part of
part of
part of
European theatre of World War I
location
part of
has parinstatopic's main category
Europe
World War I
4
Category:European theatre of World War I

Figure 2: WWI war theatre ontology.

a subpart of the WWI ontology (e.g. Finnish civil war). We make sure we had extracted all the battles related to WWI, direct or indirect. One of our findings is that WWI data on Wikidata are unbalanced. The European war theater is much more developed than other parts of the world, and even within the European war theater, the Western front is more developed than others (e.g. Balkans Campaign or Italian front). For the less developed parts, the data can be incomplete, have wrong or even missing coordinates. In some cases we fixed them and cleaned manually for our app, though we did not have resources to add complete information from other sources. Nevertheless, adding battle data on the app can be useful, since it helps users to place anchor location references into the map.

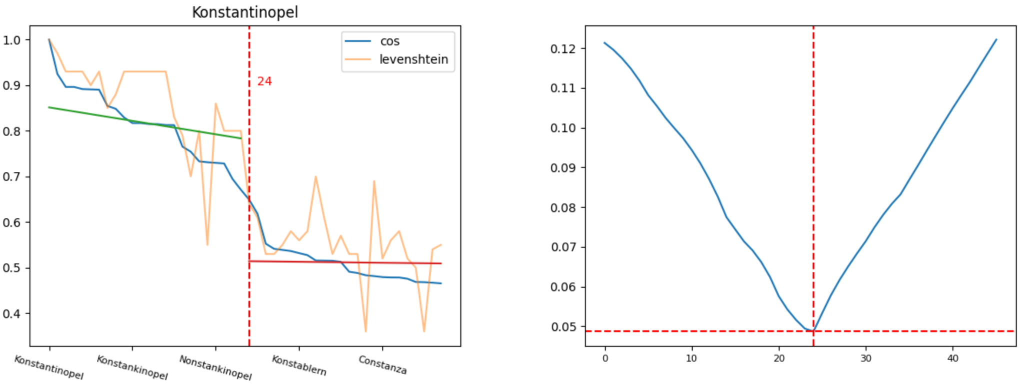
4. Data preprocessing

4.1. Entity mention normalization

In order to address the OCR-induced orthographic variation, we adopted a partial-matching-based approach [14, 16], i.e. finding candidate spellings via feature-vector similarity comparisons. According to [17], character-ngram representations provide significant improvements over word vectors for French (14%), German (28%) and Finnish (66%) monolingual retrieval in terms of *mean average precision*. Moreover, [17] establishes a high correlation (0.935) of this accuracy boost with the mean word length per language, thereby indicating its pertinence for morphologically more complex languages, e.g. Finnish and German.

Concretely, character-level n-gram counts are extracted for all unique entity tokens of a given newspaper-year domain, i.e. each entity mention is represented by an n-gram vector. In a next step, the top 100 most frequent mentions are used as query terms over the total set of unique mentions of the underlying newspaper-year domain. Given the resulting sparse vector representations, cosine distance has been employed due to its robustness w.r.t. vector dimensionality and successful application to syntactic matching problems [18].

Ultimately, a cut-off threshold needs to be determined for the most salient spelling variants in the cosine ranking w.r.t the query term. To achieve this, we apply the extended *L-method*, a fully automatic separation algorithm for large evaluation graphs [19]. The rationale behind this method lies in the assumption that evaluation graphs can be approximated by means of two



(a) Extended L method applied on the query term *Konstantinopel*

(b) Combined RMSE loss function optimal cutoff for query term *Konstantinopel.*

Figure 3: Most salient nearest neighbour retrieval via extended *L-method*

regression lines L_c and R_c separated by a cutoff point c that minimizes the linearly interpolated objective functions for both sides of the cutoff point:

$$L(c) = \frac{c-1}{n-1} RMSE(L_c) + \frac{n-c}{n-1} RMSE(R_c) \tag{1}$$

where n is the total length of the set of unique place mentions in the newspaper-year domain and RMSE stands for Root-Mean-Square Error.

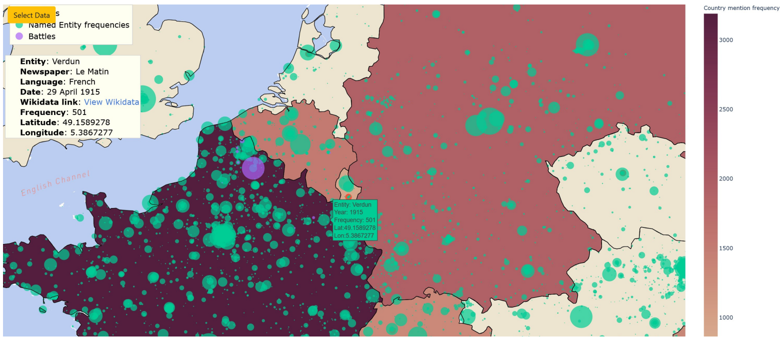
For example, Figure 3a shows name mentions ranked according to their bag-of-ngram cosine similarity to the query term *Konstantinopel.* The method successfully captures OCR-conditioned near-duplicates e.g. *Konstankinopel*, *Nonstankinopel* and distinguishes these from dissimilar terms, e.g. *Konstablern*, *Constanza.* The objective function is depicted in Figure 3b where the optimal cut-off point is visualized by the vertical dashed red lines in both plots. The optimal regression lines on the left and right of the cut-off point are highlighted in green and red respectively in Figure 3a. In our case of particularly large evaluation graphs the algorithm is applied iteratively with each computed cutoff point serving as the basis for the respectively consecutive iteration until convergence [19].

4.2. Georeference extraction

Subsequently, the normalized place mentions are mapped onto a corresponding cartographic representation. Following Yao's theoretical concept of *geocoding* as an act of *discrete georeferencing* [20], we fall back on *OpenStreetMap* data served by the *Nominatim* open-source georeferencing service[4]. Dedicated historical map services e.g. *OpenHistoricalMaps*[5] or *RunningReality* could not be used because of a lack of Application Programming Interface (API).

[4] https://nominatim.org/release-docs/develop/develop/Ranking/

[5] *OpenHistoricalMap* https://openhistoricalmap.org



Entity: Verdun
Newspaper: Le Matin
Language: French
Date: 29 April 1915
Wikidata link: View Wikidata
Frequency: 501
Latitude: 49.1589278
Longitude: 5.3867277

Entity: Verdun
Year: 1915
Frequency: 501
Lat:49.1589278
Lon:5.3867277

Figure 4: Visualizing every entity in *Le Matin* (1913-1915) through the map module

While contemporary maps do not account for historical geographical entities, such as *Austria-Hungary* or *Ottoman Empire*, we leverage the statistical properties of the dataset and the fact that major war events tend to appear in clusters around less prominent, peripheral locations in the vicinity of country borders, e.g. Western Front. We therefore argue that the eventual density and number of cartographic events serve as a proxy confidence for the geotagger output. Thus, the development and provision of such utilities remains a consideration for future work.

5. Web interface

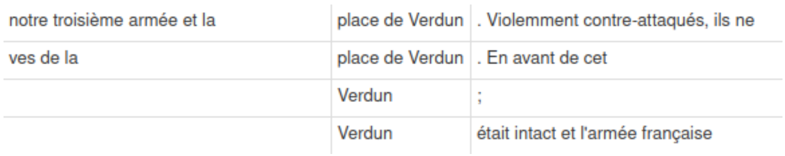
5.1. Map navigation module

We created a base map that shows the borders and the capitals of countries between 1913 and 1920. This map is based on [21] which combines multiple sources such as [22] or the *Territorial Change Dataset* by [23] in order to represent country borders from 1886 to 2019. On top of the base map the tool shows frequencies of the locations mentioned in the data subset the user has selected. The tool can show the full dataset or filter it by language, newspaper title or year.

The map is exemplified in Figure 4. The more frequent a location is in the selected dataset the bigger it appears on the map. Each country on the map is associated with a colour: the darker the colour, the more frequent that country. The map also shows contextual information such as capital cities and battles that occurred in the selected time period. The longer a battle lasted and the bigger it will appear on the map. Hovering over or clicking on an entity will show the metadata related to it such as its frequency, the newspaper it appears in or the link to the Wikidata resource associated with that entity mention if that link has been found[6].

It can be seen in Figure 4 that most named entities mentioned in *Le Matin* between 1913 and 1915 are located in France, as well as in Germany, Switzerland and Belgium. Those countries are also more frequent than other surrounding countries. This would suggest that *Le Matin* had a Eurocentric view of the war and focused on the Western front, even though the fighting spread also to Africa and Asia. This is an example of biases that are relevant for historical research and could be easily found with our interface.

[6] More details on the interface can be seen in our tutorial video:https://youtu.be/iIpEvM9IFaM



notre troisième armée et la	place de Verdun	. Violemment contre-attaqués, ils ne
ves de la	place de Verdun	. En avant de cet
	Verdun	;
	Verdun	était intact et l'armée française

Figure 5: Concordancer table view

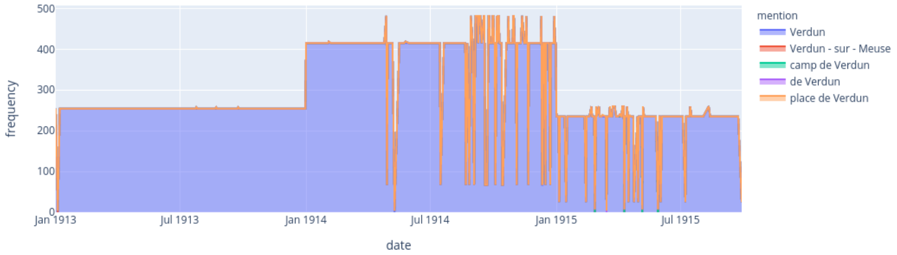


Figure 6: Concordance time series view

5.2. Concordancer module

To enable close reading functionality, we implemented a concordancer that shows every mention of a location surrounded by its left and right context. By default, these contexts are limited to 5 words before and after the mention. Each mention in the table is associated with a link that redirects the user to the NewsEye platform. There the user can extend the research by reading the article and the newspaper where an entity mention appears. The concordancer is linked to a plot that shows the distribution of entity mentions across time, by language or by newspaper.

Figure 5 shows every occurrences of ”Verdun” in *Le Matin*(1913-1915) while Figure 6 show its distribution across that period. Because of its proximity with the German border, Verdun was a key position in the conflict. Thus, it is of no surprise that Verdun is nearly exclusively mentioned in war reports or articles related to the conflict. Interestingly, other location mentions related to Verdun reinforce the relation of the city with war. For instance, Verdun is sometimes mentioned as ”place de Verdun”. ”place de” usually indicates a city square, however in this case, ”place de Verdun” always refers to the city of *Verdun* itself. There are also mentions of the military camp in Verdun (”camp de Verdun”), but only in articles published in 1913 referring to the Franco-Prussian war. Both these mentions insists on the importance of the city as a defensive position, even before the beginning of WWI. The only exception to that is the mention of ”Verdun-sur-Meuse”, which was the official name of the city between 1801 until 1970 and which is only mentioned to indicate the birthplace of a person.

6. Conclusion

We presented a publicly available web application aimed at supporting analysis of spatial imaginaries that are reflected in historical newspapers published during WWI. The data, provided by the NewsEye project, include automatically extracted location names, which we cleaned up and mapped to geographical coordinates. The interface summarizes large amounts of data and enables the comparison of constructed imaginary spaces emanating from place name collocations with the physical spatial dimensions of WWI. The combination of map, concordancer and original links to articles allows an easy switching between distant and close reading.

Our experiments also reveal the lack of machine-readable representations of historical knowledge, such as historically-aware gazetteers, historical event locations (e.g. battle places) and other information. Thus, even though WWI is exhaustively studied in historiography, resorting to community-driven resources, e.g. *Wikidata*, as part of the *WikimediaCommons*, rather than dedicated scholarly ones was largely driven by technical convenience.

We believe this problem is common for many use cases and should be addressed in the future by creating shared historical knowledge bases with open access points in order to facilitate the development of analysis tools for historical research. In addition, many resources that allow users to browse historical information online do not provide utilities for automatic reuse of the data. The development and provision of such utilities remains a consideration for future work.

Acknowledgments

This work has been supported by the European Union Horizon 2020 research and innovation programme under grant 770299 (NewsEye), and by the Région Bourgogne Franche-Comté, France, as part of the EMONTAL project.

References

[1] Ø. Eide, Media Boundaries and Conceptual Modelling: Between Texts and Maps, Palgrave Macmillan UK, 2015. URL: https://www.springer.com/de/book/9781137544575. doi:10.1057/9781137544582.

[2] J. Watkins, Spatial imaginaries research in geography: Synergies, tensions, and new directions, Geography Compass 9 (2015) 508–522.

[3] S. Davoudi, J. Crawford, R. Raynor, B. Reid, O. Sykes, D. Shaw, Spatial imaginaries: tyrannies or transformations?, Town Planning Review (2018).

[4] C. Blevins, Space, nation, and the triumph of region: A view of the world from houston, Journal of American History 101 (2014) 122–147. doi:10.1093/jahist/jau184.

[5] A. Doucet, M. Gasteiner, M. Granroth-Wilding, M. Kaiser, M. Kaukonen, R. Labahn, J.-P. Moreux, G. Muehlberger, E. Pfanzelter, M.-E. Therenty, et al., Newseye: A digital investigator for historical newspapers, in: 15th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2020, 2020.

[6] D. J. Bodenhamer, J. Corrigan, T. M. Harris (Eds.), The Spatial Humanities: GIS and the Future of Humanities Scholarship, Indiana University Press, Bloomington & Indianapolis, 2010.

[7] D. Cooper, C. Donaldson, P. Murrieta-Flores (Eds.), Literary mapping in the digital age, Digital research in the arts and humanities, first published ed., Routledge, Taylor & Francis Group, London New York, 2016.

[8] J. Strötgen, M. Gertz, P. Popov, Extraction and exploration of spatio-temporal information in documents, in: GIR, 2010.

[9] Govind, C. Alec, M. Spaniol, Elevate-live: Assessment and visualization of online news virality via entity-level analytics, in: ICWE, 2018.

[10] Govind, M. Spaniol, Elevate: A framework for entity-level event diffusion prediction into foreign language communities, Proceedings of the 2017 ACM on Web Science Conference (2017).

[11] W. Scheidel, Orbis: The Stanford Geospatial Network Model of the Roman World, SSRN Electronic Journal (2015). URL: http://www.ssrn.com/abstract=2609654. doi:10.2139/ssrn.2609654.

[12] E. Boroş, A. Hamdi, E. L. Pontes, L.-A. Cabrera-Diego, J. G. Moreno, N. Sidere, A. Doucet, Alleviating digitization errors in named entity recognition for historical documents, in: Proceedings of the 24th Conference on Computational Natural Language Learning, 2020, pp. 431–441.

[13] E. Linhares Pontes, A. Hamdi, N. Sidere, A. Doucet, Impact of ocr quality on named entity linking, in: Digital Libraries at the Crossroads of Digital Information for the Future, Springer LNCS, 2019, pp. 102–115. URL: https://doi.org/10.5281/zenodo.3529180. doi:10.5281/zenodo.3529180.

[14] S. Dumais, Improved String Matching Under Noisy Channel Conditions, in: Proceedings of CIKM 01, 2001, pp. 357–364. URL: https://www.microsoft.com/en-us/research/publication/improved-string-matching-under-noisy-channel-conditions/.

[15] P. Kantor, E. Voorhees, The trec-5 confusion track: Comparing retrieval methods for scanned text, Information Retrieval 2 (2000) 165–176. doi:10.1023/A:1009902609570.

[16] R. Jin, C. Zhai, A. Hauptmann, Information retrieval for ocr documents: A content-based probabilistic correction model, Proceedings of SPIE - The International Society for Optical Engineering 5010 (2003) 128–135. doi:10.1117/12.472838, document Recognition and Retrieval X ; Conference date: 22-01-2003 Through 24-01-2003.

[17] P. Mcnamee, J. Mayfield, Character n -gram tokenization for european language text retrieval, Information Retrieval 7 (2004) 73–97. doi:10.1023/B:INRT.0000009441.78971.be.

[18] R. S. Mishra, K. Mehta, N. Rasiwasia, Scalable approach for normalizing e-commerce text attributes (SANTA), CoRR abs/2106.09493 (2021). URL: https://arxiv.org/abs/2106.09493. arXiv:2106.09493.

[19] S. Salvador, P. Chan, Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms, in: 16th IEEE International Conference on Tools with Artificial Intelligence, 2004, pp. 576–584. doi:10.1109/ICTAI.2004.50.

[20] X. Yao, Georeferencing, geocoding, in: R. Kitchin, N. Thrift (Eds.), International Encyclopedia of Human Geography, Elsevier, Oxford, 2009, pp. 458–465. URL: https://www.sciencedirect.com/science/article/pii/B978008044910400448X. doi:https://doi.org/10.1016/B978-008044910-4.00448-X.

[21] G. Schvitz, S. Rüegger, L. Girardin, L.-E. Cederman, N. Weidmann, K. S. Gleditsch, Mapping The International System, 1886-2017: The CShapes 2.0 Dataset, Journal of Conflict Resolution (2021). URL: https://journals.sagepub.com/doi/full/10.1177/00220027211013563. doi:10.1177/00220027211013563.

[22] K. S. Gleditsch, M. D. Ward, Interstate system membership: A revised list of the independent

states since 1816, International Interactions 25 (1999) 393–413.

[23] J. Tir, P. Schafer, P. F. Diehl, G. Goertz, Territorial changes, 1816–1996: Procedures and data, Conflict Management and Peace Science 16 (1998) 89–97. URL: https://doi.org/10.1177/073889429801600105. doi:10.1177/073889429801600105. arXiv:https://doi.org/10.1177/073889429801600105.