Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

How to visualize the geography of Swiss history

André Bruggmann
Department of Geography
University of Zurich
Winterthurerstr. 190
8057 Zurich, Switzerland
andre.bruggmann@geo.uzh.ch

Sara I. Fabrikant
Department of Geography
University of Zurich
Winterthurerstr. 190
8057 Zurich, Switzerland
sara.fabrikant@geo.uzh.ch

Abstract

Efficient and effective access to and knowledge construction from massively growing spatial and non-spatial databases available online today have become major bottlenecks for the rapidly evolving information society at large. We present a geovisual analytics framework to deal with spatio-temporal knowledge extraction from rapidly growing, and increasingly massive, digital text databases largely untapped for spatio-temporal analyses. Our interdisciplinary, theory-driven approach combines text data mining methods, currently employed in GIScience and geovisual analytics, to re-organize and visualize a semi-structured online dictionary about Swiss history, made available to the general public. We automatically extract spatial, temporal, and thematic information from the text archive, and make it visually available to an information seeker interested in Swiss history, through empirically validated spatialization display techniques (e.g., network visualizations and self-organizing maps). In this case study, we specifically illustrate how spatial relationships between Swiss toponyms can be extracted, analyzed, and visualized using our proposed approach. With this interdisciplinary geovisual analytics approach situated at the nexus of digital humanities, information science, and GIScience we hope to provide new transdisciplinary solutions to facilitate information extraction of and knowledge generation from information buried in vast unstructured text archives.

Keywords: geovisual analytics, geographic information retrieval, information visualization, text mining, digital library.

1 Introduction

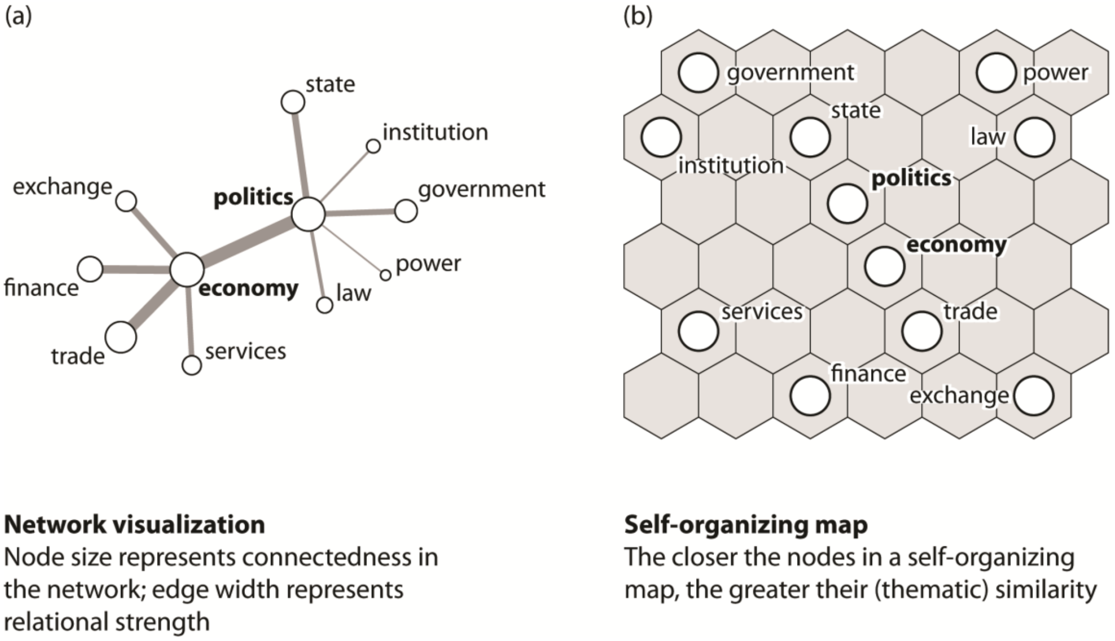
Large online digital libraries such as, the Encyclopedia Britannica or Google Books, for example, provide today's information seekers access to massive collections of unstructured digital text data and diverse multimedia content. Libraries play an important role in the humanities and the social sciences, where text documents have been central data sources for a very long time before digitization, but they are still largely untapped for spatio-temporal analyses. With massive text collections becoming available digitally, knowledge generation challenges have emerged, tackled by a variety of research communities besides GIScience (i.e., computer science, information science, digital humanities, etc.). In this context, automated text analytics techniques and tools provided by the geographic information retrieval (GIR) community coupled with powerful visuo-spatial geovisual analytics (GeoVA) interfaces seem especially relevant. GIR deals with automatically extracting relevant spatio-temporal information from digital text archives across place, over time, and on various topics and themes. GeoVA is concerned with making this information available to an information seeker through powerful, interactive graphic displays to facilitate information exploration and knowledge construction. Both GIR and GeoVA aim at revealing hidden patterns in large, typically unstructured text databases, and in doing so allow users to more easily explore emerging relationships between documents, and eventually increase sense making from vast text databases.

We present an interdisciplinary, theory-driven approach which combines GIR and GeoVA methods to re-organize and visualize semi-structured texts from a digital dictionary about Swiss history. We aim to re-organize spatial, temporal, and thematic characteristics automatically extracted from the text data archive, and make this information visually available to an information seeker interested about Swiss history. Our interdisciplinary approach may provide likeminded GIScientists new inspirations to further advance spatio-temporal methods and tools for similar kinds of datasets in a transdisciplinary context.

2 Background

A variety of automated methods have been developed in GIR to extract spatio-temporal and thematic information from digital text data sources. For example, [8] demonstrate how toponyms (e.g., London) can be automatically extracted from a massive, historical, unstructured text archive, and present novel approaches to deal with disambiguation problems (e.g., does London refer to the Capital of the U.K., or the town in Ontario, Canada?). Other toponym disambiguation solutions can be found in [7], [15] or [17]. Significant advances have also been made in automatically extracting temporal information from digital text sources. [1] and [2], for example, review current research trends in temporal information retrieval. Various date extraction tools (e.g., HeidelTime) have been developed and tested so far [24]. Similarly, attractive solutions to automatically identify thematic relations in text sources, based on the quantification of thematic similarity of texts (i.e., by using word similarity or the similarity of parts of sentences in texts), have been proposed by [23]. Their *Probabilistic Topic Models* (TM) are already widely used in the digital humanities communities. [20]

Figure 1: Network visualization (a) and self-organizing map (b).



demonstrated how the results of the TM approach may be input to graph-theoretic clustering methods, to automatically assign a thematic label (e.g., economy, politics, etc.) to grouped text documents [3].

Moreover, the visualization communities have proposed solutions to depict multivariate (i.e., spatial, temporal and thematic) numerical and non-numerical data. GeoVA has already been successfully employed to visualize uncovered relations in vast text databases. For example, [9], and [22] introduce the *spatialization framework,* which includes a systematic approach to transform high-dimensional data sets into lower-dimensional, spatial representations for facilitating data exploration using spatial metaphors. Self-organizing maps (SOM) and network maps are good examples of this (see Figure 1). Both mapping approaches depict documents that are similar in content close to one another in the visualization, as illustrated in Figure 1. Emerging themes and respective document clusters are schematically represented in Figure 1. The network map (1a) depicts text documents as nodes on a relational semantic network. Document clusters that are similar in content are connected with one another. The bigger the nodes, the stronger the connectedness of the respective document clusters with other documents in the database. The thicker the edge between nodes in the network, the stronger the thematic relationship between two corresponding document clusters in the database.

SOMs (a neural network method) project multivariate input data onto a two dimensional, topological space, typically represented by a regular tessellation (i.e., hexagons) [14]. The neurons in the SOM have the same attributes as the input data. They are placed near each other in the map if they share similar attributes, and are therefore similar in content [21]. The original data are then mapped as points onto neurons with thematically most similar attributes, thus documents of similar content cluster with respective neurons, as illustrated in Figure 1b.

The *spatialization framework* has been applied in various ways for knowledge exploration from vast multivariate (numerical) datasets, including the temporal data dimension. For example [6] applied the SOM techniques to visually explore multivariate quantitative (e.g., census) and qualitative (e.g., open-ended survey responses) data sets. The SOMs facilitated the analysis of the characteristics of survey respondents, the socio-demographic characteristics of San Diego neighborhoods, and the characteristics of the utterances respondents used to describe these neighborhoods.

3 Case Study

Based upon the empirically validated *spatialization framework* by [9] we now outline our interdisciplinary visual text analytics approach applied to the Historical Dictionary of Switzerland (HDS) [12]. We chose this particular dataset for several reasons. As a typical example of an online digital library it specifically contains spatial, temporal, and thematic information. It serves as a proto-typical secondary data resource for researchers and the general public interested in Swiss history. The multi-lingual HDS (i.e., German, French and Italian) consists of 36,188 articles related to the history of Switzerland. For our case study we chose to analyze only the German version of the HDS. The dictionary is structured by article categories, such as *thematic contributions* (e.g., events, institutions, etc.), *geographical entities* (e.g., municipalities, Cantons, etc.), *biographies*, and by articles about historically important *families*. Currently, the articles are organized in alphabetical order. There are no possibilities to query the articles according to spatial or temporal criteria.

Our visual text analytics approach includes two phases: automated text analytics and visualization. In a first step, we extract spatial, temporal, and thematic information from the HDS articles, using the well-established GIR methods mentioned above (i.e., [8]). We extract historically relevant locations (i.e., toponyms) by first identifying candidate toponyms in the HDS with the Swissnames gazetteer [25]. This gazetteer consists of 156,755 toponyms occurring on Swiss topographic maps on a scale of 1:25,000. Following that, we resolve disambiguation issues, as described in [8]. We then employ *HeidelTime*, a tool developed by [24] to extract historically relevant dates. *HeidelTime* is based on the TIMEX3 annotation standard and the markup language TimeML [18, 19]. *HeidelTime* allows to automatically retrieve dates (e.g., 07/09/2002), periods of time (e.g., 17th century), and other temporal information from texts. Finally, we employ the above-mentioned TM method [23], available in the Text Visualization Toolbox (TVT) in MATLAB [11]. This text analytics phase yields automatically retrieved information, structured in data tables including spatial, temporal, and thematic information.

In the second phase of our approach, we visualize the retrieved information. The main aim is to create an interactive proof-of-concept interface which allows users to gain new insights into the history of Switzerland, based on the re-organization of spatial, temporal, and thematic relations extracted from articles stored alphabetically in the HDS.

In the next sections we illustrate this by example, concentrating on spatial data automatically extracted from the HDS text archive. The employed GIR algorithm extracted 13,719 distinct toponyms (e.g., names of cities, municipalities, villages, historical places, and water bodies), that appear at least once in the HDS database. In total, we retrieved 169,094 toponyms. Table 1 below details the extracted Swissnames toponym categories, the total number of toponym occurrences per category, and their corresponding percentages.

Table 1: Swissnames categories and toponym occurrences.

Toponym Categories	Total	Percent
Cities, municipalities, villages	137,751	81.5
Areas (e.g., forests)	14,693	8.7
Single objects (e.g., churches, castles)	6,917	4.1
Rivers and lakes	4,972	2.9
Mountains	2,152	1.3
Valleys	1,726	1.0
Passes	544	0.3
Miscellaneous	339	0.2
Total	169,094	100

Populated places such as cities, municipalities and villages account for 81.5 percent of all toponym occurrences in the HDS. Given that "people make history", it is not surprising that this kind of spatial information is the most relevant in the HDS. The remaining 18.5% toponym occurrences are non-urban areas (e.g., forests), individual features (e.g., churches, castles, etc.), water bodies, and landforms such as, mountains, valleys, and passes, including other miscellaneous objects.

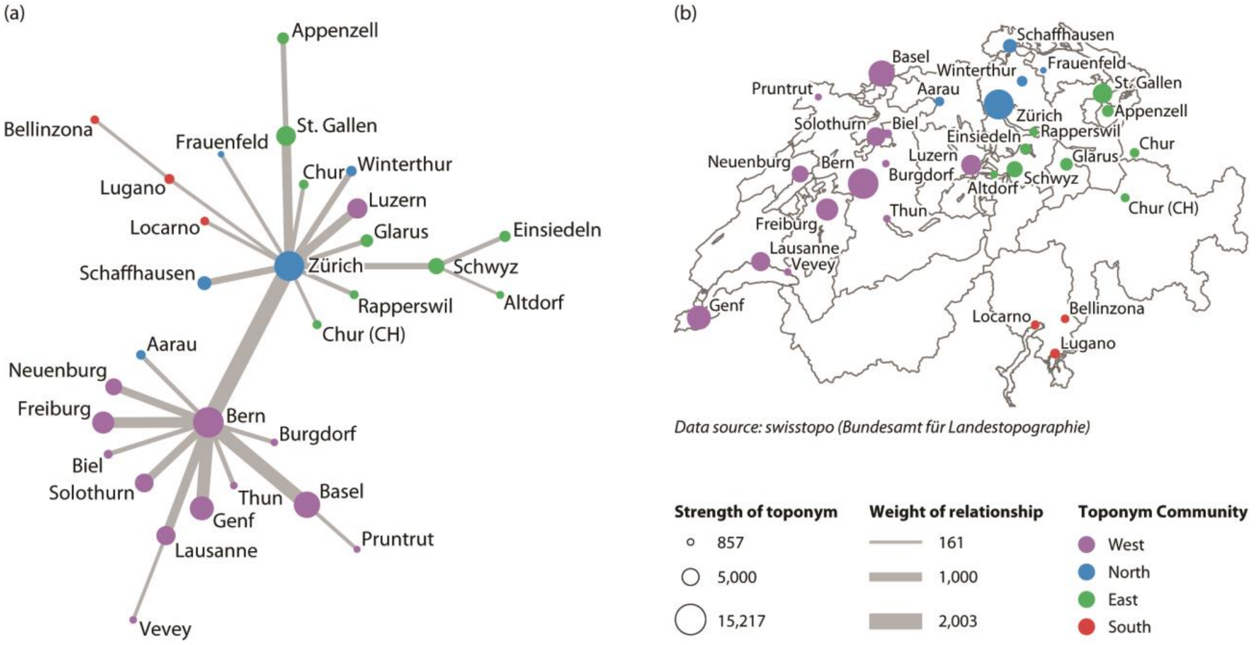
We then generate a network spatialization, similarly to Figure 1 with the extracted toponyms. Following [10]'s approach, we assume a relationship between two toponyms, if they both co-occur in the same HDS article. The overall strength of a relationship between two toponyms represents the sum of co-occurrences (divided by two) across all articles where both toponyms co-appear at least once. This weighted toponym matrix is then input to the Network Workbench (NWB) tool [16].

4 Results

Figure 2a depicts a spatialized network consisting of toponyms that, overall, occur at least 360 times in the HDS. In other words, toponyms co-appear on average once per every 100 articles. We identified 43 toponyms that meet this requirement. We excluded thirteen toponyms from this set, as they are considered stop words (e.g., *castle*). The GIR algorithm did not already exclude these. The remaining thirty toponyms account for 28.8% of the overall toponym occurrences in the HDS.

We chose the GEM layout algorithm to visualize the network, as to avoid edge crossings. We chose to only visualize the structural most important relationships using a minimum spanning tree (MST) pathfinder algorithm. We ran the Blondel community detection algorithm [3] on the weighted input matrix in NWB, to identify groups of similar toponyms which are depicted as nodes in the network (a), and in the map (b) in Figure 2. The Blondel community algorithm detects toponym groups that have strong within-group relationships, and separates toponyms that only have weak relationships. The detected communities are shown as differently colored nodes in Figure 2. We apply the visual variable line width to depict the strength (i.e., weight) of toponym relationships. The importance of a toponym in the network is shown by varying its node size. Toponym importance is computed as the sum of all weighted relationships for a toponym with all other toponyms in the network. The larger the node, the more important is the toponym in the network. Similarly, the thicker a link between toponyms on the network, the stronger their inter-relationships.

Figure 2: Toponyms relationships in text space (a) and in geographic space (b).



Data source: swisstopo (Bundesamt für Landestopographie)

Figure 2b shows the spatial distribution of the extracted toponyms on a map of Switzerland. Again, node size represents the importance of the toponym, as explained above. We can now compare and contrast the extracted toponym patterns in the network spatialization and in the map. This allows for first visual inspection of the employed methods, and further validation and evaluation. A first striking result is that the geographic pattern (2b) is replicated in the network spatialization (2a). In other words, toponyms that are close in distance in geographic space are also closely related in text co-occurrence space. Two large nodes dominate the network in Figure 2a. With *Zürich*, currently Switzerland's financial center, and *Bern*, the country's political capital, two well-known cities in Switzerland are prominently depicted in the center of the network. They are both also strongly related. These two cities form a major axis in the network linking a *West*, *North*, *East*, and *South* toponym cluster. The bottom part of the network in Figure 2a shows *Bern* as a major hub for the *West* cluster, except for *Aarau*. Considering the upper part of the network in Figure 2a, the nodes connected to *Zürich*, except for *Luzern*, are all part of the *East*, *North* and *South* clusters. *Aarau* and *Luzern* look like outlier in the network space. The map in 2b might explain this: *Aarau* is almost equally far apart from the *West* community cluster and the *North* communities. While *Luzern* is geographically located in close proximity to many nodes in the *West* cluster, it has its strongest relationships with *Zürich*.

A center-periphery pattern is visible both in the network and in the map. The smaller towns *Bellinzona* and *Einsiedeln* are not directly connected to the large center nodes *Bern* and *Zürich*. They are connected to regional center nodes like *Lugano* and *Schwyz*. This might be an indicator of local spatial clustering, and it might suggest that the network spatialization not only reveals horizontal spatial relationships, but also reproduces a spatial settlement hierarchy.

The visualization also reveals that the toponyms *Biel* and *Altdorf* are wrongly located in the map. The city of *Chur* not only exists in Switzerland, but also in the *Principality of Liechtenstein* by mistake, as the Swissnames gazetteer also contains toponyms from the *Principality of Liechtenstein*. In other words, the visual analytics approach also supports us in identifying algorithmic problems. Improvements of the algorithm will be evaluated in future work.

5 Discussion

Interestingly, geographic distance has a strong effect on the relationships between toponyms extracted from a historic text database. Indeed, extracted node clusters exhibit a highly spatially auto-correlated pattern in the map, and in the spatialized network. Already in 1971 Tobler & Wineburg [27] predicted unknown locations of historic *Cappadocian* towns using a gravity model, based on co-occurrences of place names in old Assyrian records. They speculated that interactions between cities is proportional to their populations, that is, the larger the population of two cities, the more interactions occur between them. Similarly, following "Tobler's Law" [26] they postulated that if cities are mentioned together more often on a *Cappadocian* tablet, they must be closer to one another in geographic space, compared to cities that are mentioned less often. For the data set we analyzed in this study, [27]'s speculation seems to predict very well. Geographic distances seem to be of specific importance in a historical dataset, as spatial separation was more difficult to overcome in the past than today. The concept of time-space convergence [13] which relates to the amount of space that can be covered in a given time period seems important to mention in this context. With the advancement of transportation technology people are able to cover larger distances in shorter amounts of time. In fact, the spatialized toponym network pattern reproduces the current transportation corridors in Switzerland. *Zürich* and *Bern* were then, and are still now, two major transportation hubs in Switzerland, perhaps because they are centrally located, and still today they are central nodes on a transportation network that links the Western and the Eastern parts of Switzerland. Other cities of high importance for Switzerland in the present as well as in the past such as *Basel* and *Genf* are not represented as central nodes in the network spatialization. One possible reason for this could be that relationships between these cities at the periphery might have especially in the past and also today been stronger with places located in neighboring countries than with cities in Switzerland, for instance, with places in France for *Genf* and *Basel,* and places in Germany for *Basel*. In our study, toponyms are only considered if they are located in Switzerland or in the Principality of Liechtenstein. An approach how to handle such edge effects can be found for example in [5].

6 Summary and Future Work

The aim of this study was to develop a framework based on GIR and GeoVA approaches to automatically uncover and visualize spatial, temporal, and thematic information buried in a digital dictionary about Swiss history (HDS). We showcase our approach focusing on the spatial information available in the text archive, and present a network spatialization based on co-occurrences of toponyms found in HDS articles. The uncovered network of toponyms illustrates the strong effect that geographic distance has on the historical relationships of places in Switzerland. The visual displays also helped us to uncover potential limitations of the employed GIR approach which we shall address in the future.

We are currently working towards an automated temporal analysis of the HDS articles to allow for change detection in the spatial structure and organization of toponyms in Switzerland's history. We also aim at assessing and visualizing thematic relationships (e.g., economy, politics, etc.) between toponyms in the HDS corpus, and how these might have changed over time. This may help to better explain the uncovered structure and strength of toponym relationships.

A further next step will be concerned with developing a dynamic and interactive user interface. We are currently evaluating various frameworks for online visualization including the Data Driven Documents (D3) technology [4], to complement the existing online HDS. D3, a JavaScript library, provides methods to create powerful and interactive visualization components for the Web. One component of the interface will allow users to query specific articles by space (i.e., through a cartographic map), by time (i.e., by selecting time slices), and by theme (i.e., using thematic filtering options). A second component of this interface will allow users to visually explore spatial, temporal, and thematic relationships by means of network spatializations and SOMs, as shown in Figures 1 and 2. Network spatializations might allow users to uncover hidden relationships between spatial entities over time, and, for example, how inter-city relationships might have changed over time. We envision dynamic network visualizations to emphasize *change* in the explored historical database. Thematic information will be re-organized using SOMs, which may serve as base layer onto which the change of events over time might be depicted. Spatial entities may be projected onto the self-organizing map as well, of course, and dynamically visualized, using the temporal information extracted from the text documents.

Acknowledgements

We would like to thank Curdin Derungs, Jannik Strötgen and Julian Zell who specifically helped us to implement the GIR part of our research. We are also grateful to Ross S. Purves and Damien Palacio for their invaluable feedback on this research project.

References

[1] O. Alonso, M. Gertz and R. Baeza-Yates. On the Value of Temporal Information in Information Retrieval. *SIGIR Forum*, 41(2): 35-41, 2007.

[2] O. Alonso, J. Strötgen, R. Baeza-Yates and M. Gertz. Temporal Information Retrieval: Challenges and Opportunities. In *Proceedings of the 1st International Temporal Web Analytics Workshop (TWAW 2011)*, 2011.

[3] V. D. Blondel, J.-L. Guillaume and R. Lambiotte. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008.

[4] M. Bostock, V. Ogievetsky and J. Heer. D3: Data-Driven Documents. In *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis),* 2011. http://d3js.org/ (March 2014).

[5] A. Bruggmann. Netzwerkvisualisierung der Ostschweiz. Zurich: University of Zurich, 2012.

[6] R. Burns and A. Skupin. Towards Qualitative Geovisual Analytics: A Case Study Involving Places, People, and Mediated Experience. *Cartographica,* 48(3): 157-176, 2013.

[7] D. Buscaldi. Approaches to Disambiguating Toponyms. *The SIGSPATIAL Special - Letters on Geographic Information Retrieval*, 3(2): 16-19, 2011.

[8] C. Derungs and R .S. Purves. From text to landscape: locating, identifying and mapping the use of landscape features in a Swiss Alpine corpus. *International Journal of Geographical Information Science*, 2013. DOI: 10.1080/13658816.2013.772184.

[9] S. I. Fabrikant and A. Skupin. Cognitively Plausible Information Visualization. In J. Dykes, A. M. MacEachren and M.-J. Kraak, editors, *Exploring Geovisualization*, 667-690, 2005.

[10] B. Hecht and M. Raubal. GeoSR: Geographically Explore Semantic Relations in World Knowledge. In L. Bernard, A. Friis-Christensen and H. Pundt, editors, *11th AGILE International Conference on Geographic Information Science*, 2008.

[11] S. R. Hespanha and J. P. Hespanha. Text Visualization Toolbox - a MATLAB toolbox to visualize large corpus of documents. 2011. http://www.ece.ucsb.edu/~hespanha (March 2014).

[12] Historical Dictionary of Switzerland (HDS). 2014. http://www.hls-dhs-dss.ch/ (February 2014).

[13] D. G. Janelle. Spatial Reorganization: A Model and Concept. *Annals of the Association of American Geographers*, 59 (2): 348-364, 1969.

[14] T. Kohonen. Self-organizing maps. Springer, Berlin, 2001.

[15] J. L. Leidner. Toponym Resolution in Text. Doctoral Dissertation. Edinburgh: University of Edinburgh, 2007. http://hdl.handle.net/1842/1849 (February 2014).

[16] NWB Team. Network Workbench Tool 1.0.0. 2006. http://nwb.slis.indiana.edu (March 2014).

[17] S. Overell. The Problem of Place Name Ambiguity. *The SIGSPATIAL Special - Letters on Geographic Information Retrieval*, 3(2): 12-15, 2011.

[18] J. Pustejovsky, J. Castaño, R. Ingria, R. Sauri, R. Gaizauskas, A. Setzer, G. Katz and D. Radev. TimeML: Robust Specification of Event and Temporal Expressions in Text. In *New Directions in Question Answering*, 28-34, 2003.

[19] J. Pustejovsky, P. Hanks, R. Sauri, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro and M. Lazo. The TIMEBANK corpus. In *Proceedings of corpus linguistics*, 647-656, 2003.

[20] M. M. Salvini. Spatialization von nutzergenerierten Inhalten für die explorative Analyse des globalen Städtenetzes. Doctoral Dissertation. Zurich: University of Zurich, 2012.

[21] A. Skupin and P. Agarwal. Introduction: What is a Self-Organizing Map? In P. Agarwal and A. Skupin, editors, *Self-Organising Maps: Applications in Geographic Information Science*, 1-20, 2008.

[22] A. Skupin and S. I. Fabrikant. Spatialization. In J. P. Wilson and A. S. Fotheringham, editors, *The Handbook of Geographic Information Science*, 61-79, 2007.

[23] M. Steyvers and T. Griffiths. Probabilistic Topic Models. In T. K. Landauer, D. S. McNamara, S. Dennis and W. Kintsch, editors, *Handbook of Latent Semantic Analysis*, 2007.

[24] J. Strötgen and M. Gertz. Multilingual and Cross-domain Temporal Tagging. *Language Resources and Evaluation*, 47(2): 269-298, 2013.

[25] swisstopo. SwissNames. 2014. http://www.swisstopo.admin.ch/internet/swisstopo/de/home/products/landscape/toponymy.html (February 2014).

[26] W. Tobler. A Computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(2): 234-240, 1970.

[27] W. Tobler and S. Wineburg. A Cappadocian Speculation. *Nature*, 231: 39-41, 1971.