HOW CAN GEOGRAPHIC INFORMATION IN TEXT DOCUMENTS BE VISUALIZED TO SUPPORT INFORMATION EXPLORATION IN THE HUMANITIES?

ANDRÉ BRUGGMANN, SARA I. FABRIKANT AND ROSS S. PURVES

Abstract *Finding and selecting interesting and relevant information in large online digital text archives can be challenging. We tackle this information access problem from a geographic information science perspective using a case study exploring a semi-structured historical encyclopedia. We propose a three-pronged approach for this, based around (1) automatic retrieval of spatio-temporal and thematic information from digital text documents; (2) transformation of the extracted information to spatialize and visualize spatio-temporal and thematic structures; and (3) integration of the spatialized displays in an interactive web interface driven by a user-centered design and evaluation approach. We implemented an interactive spatialized network display to allow identification of spatio-temporal relationships hidden in the text archive, complemented by an interactive self-organizing map display to visualize thematic relationships in these text documents. We evaluated the utility and usability of the developed interface in a user study with digital humanities scholars. Empirical results show that the developed interface supports target users in the humanities uncovering latent spatio-temporal and thematic relationships and generated new insights through the spatialized text collection. Adopting this approach, we illustrate one avenue to addressing the information access problem in the digital humanities from a GIScience perspective.*

International Journal of Humanities and Arts Computing 14.1-2 (2020): 98–118
DOI: 10.3366/ijhac.2020.0247

www.euppublishing.com/ijhac

Keywords: geographic information visualization, spatialization, geovisual analytics, digital humanities, interactive information visualization

INTRODUCTION

Given suggestions that up to 80% of information stored in rapidly increasing online data archives has a reference to space or geography (e.g., Adams and Gahegan, 2016), it seems that spatial information is important for the digital information society at large. Information seekers often use spatial information including place names in their queries to access online content (Jones et al., 2008). Ballatore and colleagues (2015) contend that spatial, temporal, and thematic information always interact in information search processes, and thus argue for information retrieval (IR) systems including combined search for spatio-temporal and thematic information. They also emphasize the need to systematically evaluate exploratory search in the context of unstructured data with users.

The (digital) humanities have started to develop information-seeking systems that facilitate access to large unstructured and semi-structured text archives (Kaplan, 2015). This is not only because text documents have always played a central role in the humanities, long before digitization, but precisely because texts are increasingly becoming available in digital form through large-scale digitization efforts (e.g., Google Books), user-generated content (i.e., Wikipedia), and the social media (i.e., Twitter). Effective and efficient access mechanisms allowing information seekers in the humanities to search for interesting, meaningful, and relevant information, and which present information in relevant forms to scholars allowing exploration, rather than simply identification of relevant documents for close reading, remain rare. From within the (digital) humanities, Moretti (2005) recommended a *distant reading* approach for information search and exploration of text collections. Leveraging information visualization in the humanities, the distant reading approach aims to depict overall structures and relationships extracted from documents in a text archive, instead of simply serving humanities researchers with a series of individual (digital) text documents. He goes on to suggest graphs, maps, and trees as specific visualization means for such a *distant reading* approach in the digital humanities. Information visualizations of text archives are intended to support new insights about relationships and patterns that are explicitly and implicitly stored in the text repositories. Extending Moretti's (2005) text archive visualization suggestions, Jockers (2013) recommends combining *distant* and *close reading* views to facilitate the understanding of the individual text documents, by jointly analyzing the semantic context in which they exist.

Below we detail one solution to the abovementioned research gaps. We aim to serve both, the information retrieval (IR) community, by considering combined

spatio-temporal and thematic IR systems, including the evaluation of these systems, and most importantly, the digital humanities community by combining *distant* and *close reading* approaches in a graphic user interface aimed to explore large unstructured text archives. We bridge these gaps from a geographic information science (GIScience) perspective, and propose a sequence of three steps in our methodology: 1) the automatic retrieval of spatio-temporal and thematic information from an unstructured or semi-structured digital text archive; 2) the transformation of the retrieved information and subsequent depiction of uncovered structures in spatialized views; and 3) the integration of these spatialized views into a graphic proof-of-concept user interface to explore uncovered structures, following a user-centered design and evaluation approach. We refer readers for details about the first step to Bruggmann and Fabrikant (2016). In this paper, we will focus on steps two and three, driven by the following research questions:

- How can we visualize spatio-temporal and thematic structures and relationships extracted from unstructured and semi-structured text document collections in the humanities?
- How can we make information about space, time, and theme from unstructured and semi-structured text archives available to information seekers in the humanities, to support sense-making and the generation of new insights about these text archives?

RELATED WORK

The review of related work is organized in three themes, based on research from fields that informed our proposed three-pronged approach, leading to an interactive user-centered interface for the exploration of spatio-temporal and thematic information in humanities text archives as the targeted outcome: geographic information retrieval, spatialization research, and geovisual analytics.

Geographic information retrieval (GIR) focuses on the retrieval of thematic information related to locations, often in the form of triples of place names, spatial relationships and themes (Purves et al., 2018, Jones and Purves, 2008). For example, Derungs and Purves (2014) showcase an algorithm that automatically retrieves toponyms (e.g., cities, villages, rivers, mountains, etc.) from unstructured text documents. Their approach suggests a consistent framework, including the detection, the disambiguation, and the indexing of toponyms, leveraging a gazetteer-based approach (i.e., list of potential place names). From the field of automatic temporal information retrieval, we used *Heideltime*, a cross-domain and multilingual temporal tagger for unstructured text data (Strötgen and Gertz, 2013). *Heideltime* automatically

detects, disambiguates, and annotates temporal references (e.g., dates, periods, etc.) in text documents. Probabilistic topic modeling (PTM) is a commonly used approach to automatically retrieve thematic information or topics from text documents (Steyvers and Griffiths, 2007). For example, PTM can automatically suggest latent semantic topics (e.g., economy, politics, sports, etc.) from unstructured text documents such as Wikipedia (Salvini and Fabrikant, 2016). To do so it identifies latent topics, based on co-occurrences of words in texts, and then models individual text documents as probability distributions over identified topics. The novelty of our approach is not in the use of these individual dimensions, but rather in their combination and application to spatially and temporally rich multi-faceted historical dictionary.

To visually summarize our results, we propose spatialization; a systematic approach to transform, reorganize and visualize high dimensional spatial and non-spatial information automatically retrieved from unstructured or semi-structured text documents in lower dimensional spatialized views (Fabrikant et al., 2010). Spatialization aims to generalize and transform complex multidimensional data (e.g., spatio-temporal and thematic metadata), and to visualize uncovered structures, interconnections, and relationships in human accessible visuo-spatial displays, including graphs, maps, trees, and other visualization types (Fabrikant and Skupin, 2005), akin to Moretti's (2005) *distant reading* approach. Spatialized displays are typically built using spatial metaphors such as the distance-similarity metaphor (Fabrikant and Skupin, 2005), that is, by placing items of interest (i.e., text documents) sharing similar attributes closer to one another in a graphic display than dissimilar items. Spatialized displays are not new to the (digital) humanities. They are, for example, used to visualize social networks extracted from various document collections (e.g., Bingenheimer et al., 2011). Systematic and theory-driven methods that demonstrate the meaningful use of the spatialization framework to depict spatio-temporal and thematic information and relationships retrieved from unstructured or semi-structured digital text archives in the humanities, however, are still rare.

Finally, geovisual analytics, the science of analytical reasoning with spatial information facilitated by interactive visual interfaces (Robinson, 2017) is considered relevant for the third step in our research approach. We developed a geovisual analytics proof-of-concept web-interface to present high-dimensional and complex text-data to target users by linking lower-dimensional spatialized views in an interactive and easy-to-access, and easy-to-use graphical user interface. Geovisual analytics provides not only suggestions for creating interactive visual interfaces, it also promotes systematic user-centered design and evaluation methods (Andrienko et al., 2007). Involving target users early on in the interface design process is key to creating useful and usable user interfaces (Lewis and Rieman, 1993). Geovisual analytics has already been

applied in the humanities, for example, to interactively visualize co-occurrences of place names and commodities (Hinrichs et al., 2015). We extend prior work by combining spatio-temporal and thematic information automatically retrieved from unstructured or semi-structured text archives in the humanities, to display retrieved information for interactive data exploration. In summary, the main goal of our approach is combining the *distant* and *close reading* concepts suggested by Moretti (2005).

METHODS

We start the entire process, as shown in Figure 1, by introducing the data source of our case study (Step 0 in Figure 1). Next, we illustrate the process from the retrieval of relevant geographic information (Steps 1–2 in Figure 1) to the visualization of the extracted information in interactive displays (Steps 3–4 in Figure 1) that are assembled in the proof-of-concept graphical user interface (Steps 5–9 in Figure 1).

Data Source

We illustrate our approach with the Historical Dictionary of Switzerland (HDS), an example of a semi-structured, online text archive in the humanities (HDS). The HDS consists of 36,188 articles written by historians describing the history of the territory of Switzerland (step 0 in Figure 1). The articles cover time periods from the *Paleolithic* period (about 1.5 million years ago) until today. They are grouped into four categories: *thematic contributions* (e.g., historical phenomena and terms, institutions, companies, etc.), *geographical entities* (e.g., municipalities, mountains, rivers, etc.), *biographies*, and articles about *families* that have been important for Swiss history. HDS articles thus contain rich spatio-temporal information and cover a variety of topics. Spatio-temporal and thematic information in the HDS has not been extracted, systematically analyzed, and depicted to date. The current online version of the HDS (http://www.hls-dhs-dss.ch) offers limited querying options, including title or full text search, and article category filtering. HDS Articles cannot be directly accessed by spatio-temporal or thematic filtering methods. In Moretti's (2005) terms, interaction with the HDS is limited to simple text search functionality and *close reading*, whereas *distant reading* functionalities, giving overviews of large numbers of articles, are not provided to information seekers.

Although HDS articles are available in all four languages spoken in Switzerland (i.e., German, French, Italian, and Rhaeto-Romanic), we only considered the German version of the HDS for our case study. We thus selected German-speaking participants for our user studies to minimize translation and comprehension issues in subsequent user evaluations.

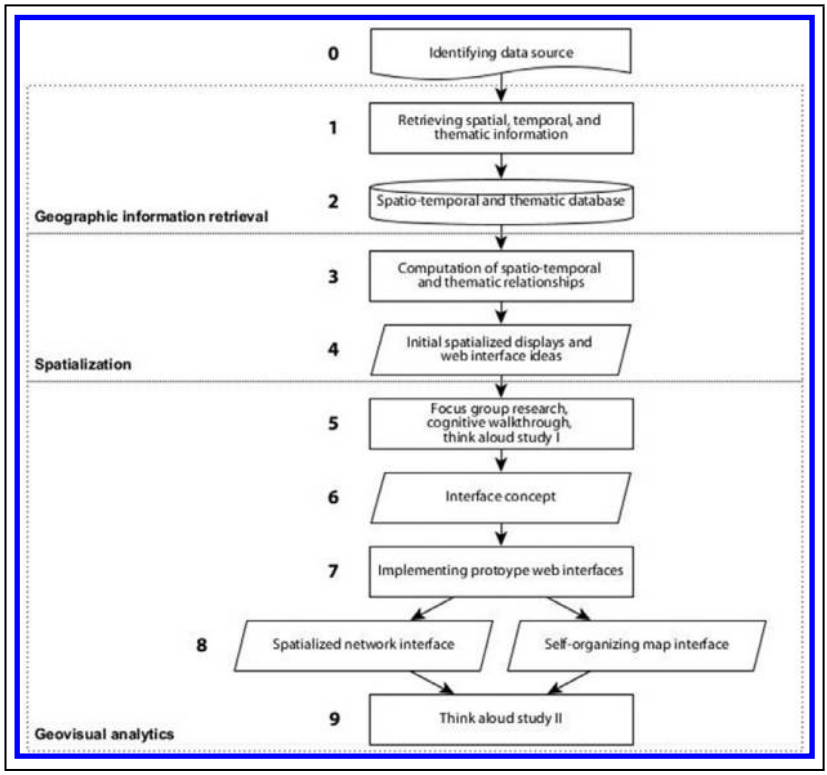

Figure 1. The overall workflow of this project, based on I) geographic information retrieval, II) spatialization, and III) geovisual analytics approaches and techniques. The numbered chart features are referenced in the text.

Geographic Information Retrieval

The geographic information retrieval steps have already been presented elsewhere (i.e., Bruggmann and Fabrikant, 2016), and are thus only briefly covered here. We extracted 322,179 Swiss toponyms (e.g., cities, towns, etc.) from the 36,188 HDS articles, of which 16,489 toponyms are unique. Next, we retrieved temporal information from articles with *Heideltime* (Strötgen and Gertz, 2013). In total, we retrieved 499,258 temporal expressions (e.g., dates, periods of time, etc.) from the HDS articles. Finally, we automatically clustered and assigned themes to a subset of 3,067 *theme* articles in the HDS, based on their topical content by means of a probabilistic topic modeling algorithm (PTM). We chose this subset of HDS articles, as they cover salient topics in Swiss history, and thus contain a great deal of thematic information which can be used to access the HDS from a thematic perspective. To find an appropriate

number of topics with PTM we first computed model fit measures (i.e., log likelihood per token) for different numbers of topics (Steyvers and Griffiths, 2007). Then, we visualized PTMs in self-organizing maps (see next section) that revealed high model fit. We qualitatively assessed generated topics by visual inspection. Both evaluation procedures are detailed in Bruggmann (2017). Based on these evaluations, we decided for a 30-topics PTM solution. The computation of the PTM yielded an article-topic matrix which assigns a vector to each article, containing a probability distribution across the 30 automatically identified topics. This article-topic matrix was then used to cluster the HDS articles, based on cosine similarity of article vectors. Articles with a similar probability distribution are more likely to be clustered to the same topic than articles with dissimilar probability distributions. We applied the community detection algorithm developed by Blondel and colleagues (2008) to cluster HDS articles. The main reason for choosing this algorithm is that the optimal number of clusters is automatically determined, compared to other common clustering methods (e.g., k-means) which require that the number of clusters be determined a priori. Hence, the 3,067 articles were automatically categorized into 28 themes. The produced information then served as an input for the subsequent spatialization process (Steps 3 and 4 in Figure 1), that is, to re-arrange and visualize uncovered spatio-temporal and thematic relationships.

Spatialization

We explored at the outset, together with digital humanities scholars invited to a focus group (see more details in subsequent sections), possible distant reading approaches in the context of the HDS. Typical questions that arose with respect to visualizing toponym relationships over time were, for example, being able to compare the strength of toponym relationships over time, across spatial scales, and across article categories, or being able to compare the membership categories of a toponym across time periods. For this we chose networks to spatialize and visualize uncovered relationships between places mentioned in the HDS. Networks explicitly symbolize linkages between items. It is also possible to model and visualize vertical (i.e., hierarchy of documents and themes) and horizontal (i.e., centrality of items) relationships in the HDS database. We selected the self-organizing map (SOM) display method to depict the thematic information we retrieved from the HDS. SOMs facilitate visual apprehension of themes and thematic clusters of HDS articles; leveraging the distance-similarity metaphor by means of visualizing semantically similar content (Skupin and Agarwal, 2008). Both, the network views and the SOM displays, assume that the more similar data items are, the closer they are located to one another.

To include a temporal dimension in our displays we computed a combined spatio-temporal co-occurrence score to generate spatialized networks over time.

Inspired by Hecht and Raubal (2008) and Salvini and Fabrikant (2016), we consider the co-occurrence score of two toponyms to be high, if toponyms co-occur often in the same HDS articles. We aggregated temporal expressions (e.g., dates, periods, etc.) to the temporal granularity of centuries. We computed temporal weights to separate weighted toponym relationships across individual centuries. For instance, given two toponyms, we consider a higher spatio-temporal co-occurrence score for these two toponyms for the 20^{th} century, if they co-occur in HDS articles that contain a high percentage of temporal expressions about the 20^{th} century. We assign a vector with temporal weights to represent the relative frequencies of temporal expressions about specific centuries in an HDS article (e.g., 18^{th} century $= 0.1$, 19^{th} century $= 0.2$, 20^{th} century $= 0.7$). We computed the co-occurrences of the 203 most frequent toponyms in HDS articles for the 18^{th}, 19^{th}, and the 20^{th} centuries, at two different spatial scales (e.g., Switzerland and the Canton of Zurich). We then analyzed and visualized the spatio-temporal relationships for each century at both spatial scales.

The visualization of spatialized networks follows Salvini and Fabrikant (2016)'s empirically validated spatialization and visualization approach, based on Montello, Fabrikant, Ruocco, and Middleton (2003), who found it to be intuitive to understand and to use. Toponyms, which co-occur frequently in the HDS articles are represented by nodes which are connected by an edge in the network (Figure 3). Toponym clusters created based on the co-occurrence of toponyms in the HDS articles are depicted by color hue. The more toponyms co-occur in the same HDS articles, the higher the probability that they are assigned to the same cluster. To compute toponym clusters, we used the community detection algorithm developed by Blondel et al. (2008).

The size of the nodes in the network represents toponym centrality using the sum of spatio-temporal co-occurrence scores to all toponyms in the network. The darker the color and the thicker the edge, the stronger the relationship between toponyms in the network. We only visualize the structurally most relevant relationships, identified with the pathfinder network scaling algorithm (Dearholt and Schvaneveldt, 1990), as suggested by Salvini and Fabrikant (2016).

Scholars invited to focus groups found it useful to see an overview of the available historic themes in the HDS, and to see individual articles embedded within those themes. So, a typical search scenario would entail identifying articles about a specific topic and/or find thematically similar articles about a specific topic. Being able to identify toponyms that are most relevant to a specific topic was also seen as useful for distant reading. We thus constructed self-organizing maps (SOMs) using the article-topic matrix as input to visualize thematic structures and interconnections in the HDS (i.e., Step 4 in Figure 1). As the empirically evaluated distance-similarity metaphor would predict (Fabrikant et al., 2006), HDS articles that are very similar in their thematic content are placed within the same hexagon (i.e., neuron) in the SOM (Skupin and

Figure 2. HDS articles belonging to the *press* (yellow) and *religion* (violet) topics. Randomly selected articles are labeled (Bruggmann, 2017).

Agarwal, 2008). Construction details on the SOM generation presented in this paper are provided in Bruggmann (2017). A small portion of the HDS SOM is visualized in Figure 2. Individual HDS articles are depicted as points in the SOM and color hues denote the thematic clusters that have been automatically assigned to the HDS articles as described in the GIR section above. Figure 2 shows some of the labels for randomly selected HDS articles belonging to two different thematic clusters: *media/press* (yellow dots) and *religion* (violet dots), separated by a thick grey boundary line.

User-centered Geovisual Analytics

The spatialized network and the SOM views were incorporated into an interactive, online web-based, user interface (Steps 4–8 in Figure 1). The implementation of the interactive web interface is the targeted outcome of our case study. We chose an iterative and user-centered interface design and evaluation approach as, e.g., Lewis and Rieman (1993), and exemplified by Roth et al. (2015), evaluated with target users in a series of empirical studies (Bruggmann and Fabrikant, 2016). We conducted three user studies (Step 5 in Figure 1). Target users of our interface were historians, interested in new media types and methods in history, and keen to experiment with easy-to-use interactive graphical interfaces to explore information in and of the humanities in general. The results of the different user studies incrementally guided the design of the graphical user interface concept and subsequently of the implemented proof-of-concept web interface (Step 6 in Figure 1). Below, we briefly summarize the empirical evaluation of our web interface. Further methodological details are provided in Bruggmann and Fabrikant (2016).

The first focus group session with five representative members provided first qualitative insights into little known needs and requirements of our target users, at a very early stage of the interface design process. Armed with these new insights, we developed paper mockups of the web interface. Next, we evaluated the paper mockups in an empirical study without users by predicting their behavior by means of a cognitive walkthrough (Lewis and Rieman, 1993). We

were thus able to already identify and remove potential interface design issues, before evaluating the interface mockups with target users. Based on the results of the cognitive walkthrough, we revised the paper mockups and evaluated them with five participants from the chosen target group, following the same procedure as with the cognitive walkthrough. At this stage, we invited participants who have had first experiences with interactive web interfaces but had not participated in the prior focus group session. This was beneficial, as we identified additional interface design issues. We also received further feedback on our interface drafts including suggestions for new functionalities, before implementing the online, web-based prototype. Based on the outcome of these empirical evaluations, we further adapted the interface concept (Step 6 in Figure 1). This iteration yielded important feedback for the implementation of the web prototype (step 7 in Figure 1).

We then developed the web-based proof-of-concept interface (Step 8 in Figure 1). The spatialized network interface was implemented using a combination of HTML, SVG, CSS, and JavaScript. The self-organizing map interface was developed using ArcGIS Online (Esri, 2017). Both displays are designed according to the visual information-seeking mantra by Shneiderman (1996): overview first, zoom in and filter, and finally, details on demand. This information visualization concept maps nicely onto Moretti's (2005) *distant* (i.e., overview first) and *close reading* (i.e., details-on-demand) concepts.

We again evaluated the web prototype in a think aloud study with target users, the final step (9 in Figure 1) in our proposed approach. Contrary to the first think aloud study, we focused not only on the usability of the interface, but more importantly, on the utility for information access, as suggested by Roth et al. (2015). For this, we re-invited all five participants who had already taken part in the focus group session at the very beginning of our case study, as they were already familiar with the project, and were interested in seeing and using the implemented prototype. We provided participants with one information access task for each of the developed linked views in the interface. We instructed them to comment on potential insights gained while using the interface. We also asked for comments on their actions and decisions, as well as potential issues they might have faced while working with the interface, as recommended by Lewis and Rieman (1993). Participants solved the tasks in individual sessions, using a mouse and a keyboard to interact with the displays shown on a computer screen in a research lab at the University of Zurich and were also asked to fill out a system usability scale (SUS) questionnaire (Brooke, 1996). The SUS is measured on a five point Likert scale, and assesses the global usability of a system based on ten questions regarding the effectiveness, the efficiency, and users' satisfaction with the system (Brooke, 1996). Participants were audiotaped during the sessions, and their interactions with the interfaces were recorded.

RESULTS

We first present the interface components of the developed web prototype (Step 8 in Figure 1) and illustrate how the iterative user-centered design and evaluation approach influenced the design of the final version of the developed web interface. Then, we focus on the results of the second think aloud study (Step 9 in Figure 1). Detailed results of Step 5 in Figure 1 are available in Bruggmann and Fabrikant (2016). Both prototype implementations are available online at: https://www.geo.uzh.ch/~abruggma (only in German).

Spatialized Network Interface

The web-interface (Figure 3) shows data at the country scale (i.e., entire area of Switzerland) in the 19th century. Focus group participants (Step 5 in Figure 1) requested network visualizations at different spatial scales, and at various temporal levels of detail. We thus implemented 1) a drop-down menu in the web interface, so that they could select a desired spatial scale, and 2) a time slider, to select the temporal level of detail (No. 1 in Figure 3). Focus group participants suggested displaying the three strongest relationships (i.e., highest spatio-temporal co-occurrence score) of toponyms not only in the spatialized network display, but also in a linked, cartographic view.

The interface provides mouseover functionality to highlight toponyms in the network display (No. 2 in Figure 3: Toponym *Zürich* is selected by mouseover), and this highlights the strongest three relationships of the selected toponym in the network display and in the linked cartographic display (No. 3 in Figure 3). Toponyms are dynamically labeled in the network display. Edges in the network can be selected by mouseover, and its respective relationships are highlighted and labeled in the network and linked map display, as suggested in the think aloud study. Participants in the think aloud study requested that titles of HDS articles that contribute most to a shown toponym relationship (i.e., with a high spatio-temporal co-occurrence score) should be displayed and hyperlinked directly with the original articles on the HDS website (i.e., *close reading* functionality). The highlighted relationships of *Zürich* (No. 2 in Figure 3) are also summarized through hyperlinks in the information window (No. 4 in Figure 3). Hyperlinks to the highest scoring HDS articles are also displayed in the information window when a user clicks on an edge in the network visualization.

Interface elements provide users with additional information (No. 5–8 in Figure 3). A button allows labeling and highlighting of the most central toponyms in the network and in the linked map (No. 5 in Figure 3). A legend helps with interpretation (No. 6 in Figure 3). Pop-up windows show additional information about the legend items. The blue information button (No. 8 in

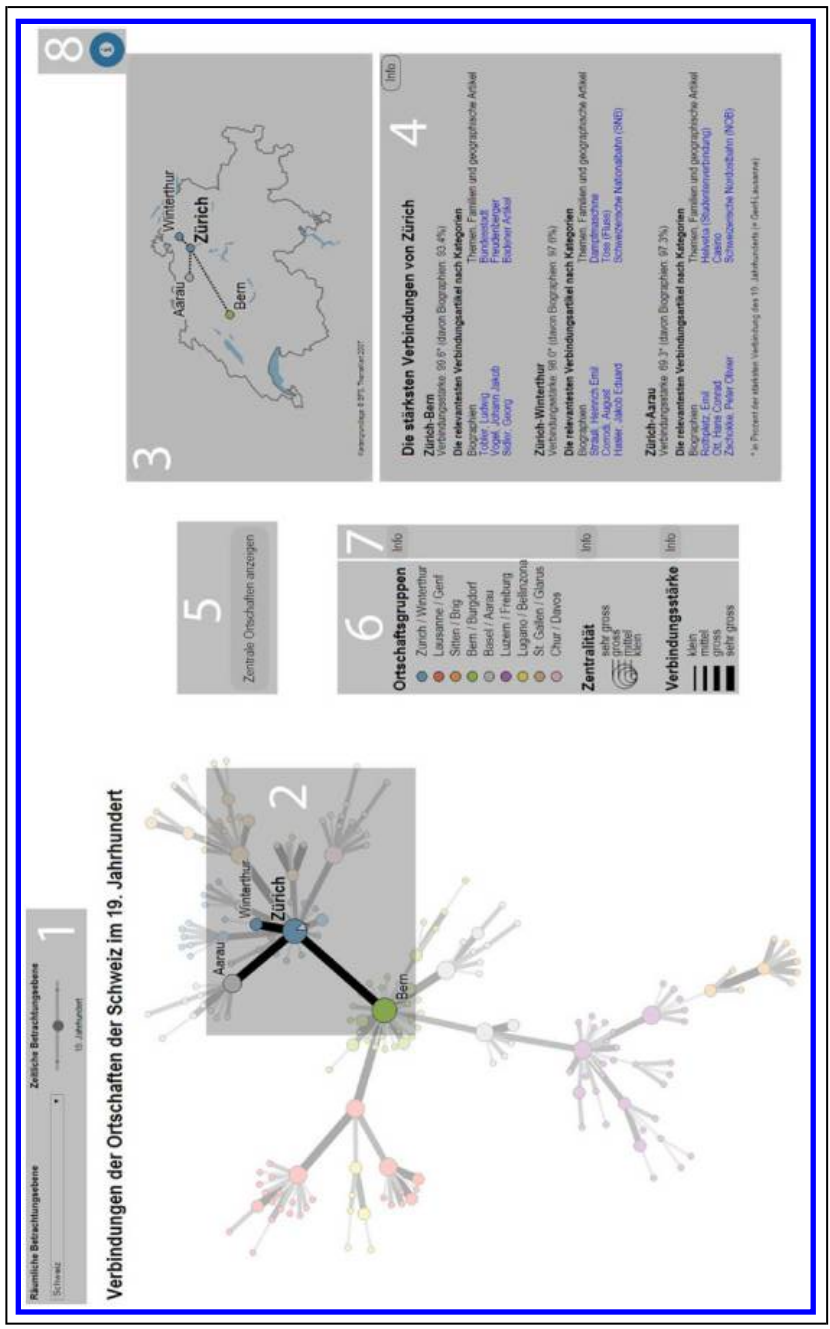

Figure 3. Spatialized network interface (Bruggmann, 2017).

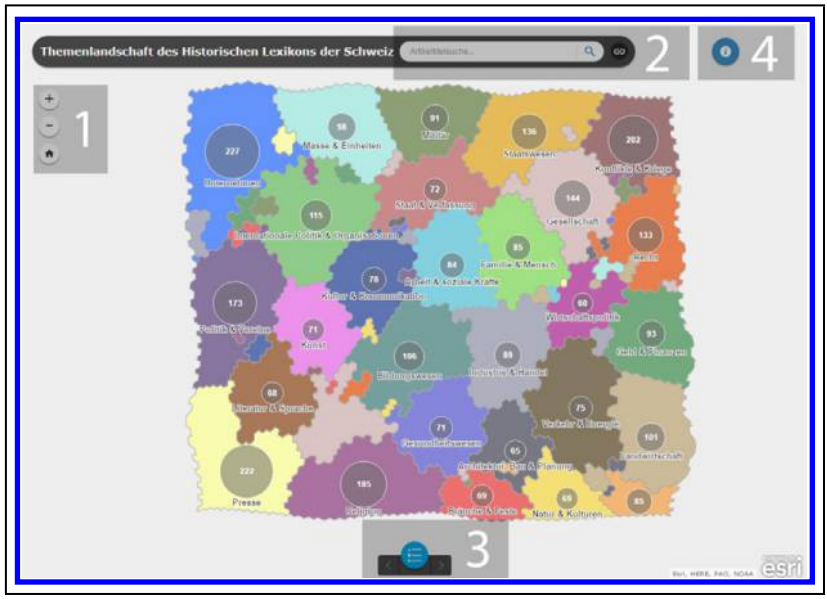

Figure 4. Overview level of the SOM interface (Bruggmann, 2017).

Figure 3) provides context information, including a hyperlink to a website with background information about the project, the used data source, the applied methods, as well as to publications and the relevant literature. This context information was made available specifically as a response to the expressed wish of focus group participants to better understand the applied methods, and algorithms used to create the network visualization, and thus, to increase trust in the displayed data. Focus group participants had wished to be able to assess the validity and meaning of the displayed information for themselves.

Self-Organizing Map Interface

The SOM interface also follows the information visualization mantra and Moretti's (2005) distant and close reading principles. It thus provides two semantic zoom levels, an overview first (Figure 4), and a detail-on-demand view (Figure 5).

Instead of displaying individual locations of HDS articles in the SOM (see Figure 2), only thematic regions are displayed at the overview level. Thematic regions in the SOM were created by resolving borders between thematically similar articles that are assigned to the same theme (shown in the same color in Figure 2). The number of articles assigned to a theme is represented by the

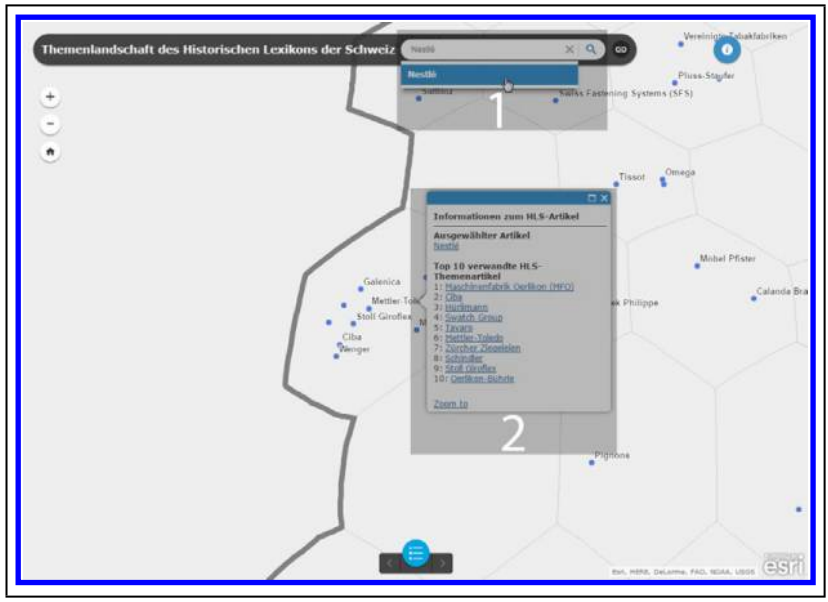

Figure 5. Detail-on-demand (i.e., close reading) view of the self-organizing map interface (Bruggmann, 2017).

size of the graduated circles in Figure 4. A zoom tool (No. 1 in Figure 4) allows switching between overview and detailed views. A text search box with an auto-complete function (No. 2 in Figure 4), a dynamic map legend (No. 3 in Figure 4), and a blue info button (No. 4 in Figure 4) are also available at the overview level. As for the spatialized network interface above, the blue information button leads to additional information about the project. After zooming into the display, a user can select individual articles directly, either by clicking on their point symbol in the map, or by typing a term into the search box (No. 1 in Figure 5). In Figure 5, a subset of articles is labeled and the HDS article *Nestlé* is selected. A pop-up window shows the title of the selected article including the titles of the ten thematically most similar HDS articles (No. 2 in Figure 5). Similarity is determined by computing the cosine between article vectors containing term co-occurrences in the article-topic matrix. All titles are hyperlinked, thus enabling direct access to the original articles on the HDS website (i.e., *close reading* functionality).

Having illustrated the functionality of the developed web interface, we now turn to the user study, aimed at evaluating the utility and the usability of the developed proof-of-concept with target users.

Second Think Aloud Study

We assessed the utility of the spatialized network interface by evaluating gained insights by participants, during a free exploration task. For that, we formulated an open-ended task, asking participants to explore the spatialized network interface, following their own interests, within a given time limit of 40 minutes. We asked participants to comment on display interactions, and about potentially gained new insights (if any), and if they did, how they had gained these insights. We videotaped user interactions with the interface, and recorded participants' utterances during exploration sessions for subsequent analysis. We found that participants used a variety of interaction components in, and information access methods with, the spatialized network display to arrive at complex insights. Following North (2006), we define insights that require interpreting and comparing diverse pieces of spatio-temporal information as complex. For example, several participants chose to explore how the centrality of a toponym (e.g., Zürich) evolved over time (e.g., from the 18th to the 20th century). This was surprising to us, as we did not expect participants to generate such deep insights within the first 40 minutes of using a novel tool. We also found that participants used both the spatialized network display (i.e., overview-first or *distant reading*) and various individual HDS articles (i.e., details-on-demand or *close reading*) to gain new insights into Swiss history from a geographic perspective. A common strategy was to first, select interesting toponym relationships in the spatialized network display, second, go through the list of most relevant and hyperlinked HDS article titles highlighted in the information window, and third, read some of the linked HDS articles directly on the HDS website. We interpret this sequence as implying that participants wished to compare the possible meanings of links between toponyms in the network display (e.g., based on assigned themes) with their own understanding after reading related HDS articles. Hence participants followed Shneiderman's (1996) information-seeking mantra: i.e., overview first, zoom and filter, then details-on-demand. Of course, this is not surprising as we designed our interface according to these principles. Participants were particularly interested in toponyms that are connected and placed in the same region of the network display, i.e., with a high spatio-temporal co-occurrence score, but that are not necessarily close in geographic space. Participants discovered that many toponyms were connected because of religion. This is not surprising, as events related to the topic of religion are important and prominent in Swiss history and thus well represented in the HDS. However, participants could confirm this fact using the spatialized views.

We evaluated the utility of the SOM display separately, because we were particularly interested in how well participants intuitively understood the visualized distance-similarity metaphor, i.e., that articles found in close

proximity to one another in the map, would be similar in content. We thus asked participants to specifically search for five HDS articles containing information across two neighboring themes in the self-organizing map (i.e., *religion* vs. *customs and festivals*) during a given time limit of 15 minutes. We expected that participants would search for articles in the border region of the neighboring themes *religion* (i.e., the purple thematic region labeled *Religion* in Figure 4) and in *customs and festivals* (i.e., the red thematic region labeled *Bräuche & Feste* in Figure 4). Of the twenty-five articles that participants identified as best-fitting to the topic "*religious customs and festivals"*, twenty-four are indeed located in this border region. Participants followed a similar exploration strategy to discover relevant articles. They firstly analyzed themes at the overview level (i.e., overview first or *distant reading*), before zooming into the border region of the *religion* and *customs and festivals* themes (i.e., zoom or filter). Finally, participants opened pop-up windows of various HDS articles in the chosen area of the map, to read original articles available on the HDS website (i.e., details on demand or *close reading*).

We also ran a system usability score (SUS) evaluation (Brooke, 1996) on both interface components to assess the general usability of the implemented interface. SUS scores range from a minimum of zero to a maximum of 100 points. The higher the score, the better the usability. Bangor, Kortum, and Miller (2008) suggest 70 for a system to be of acceptable usability, and in the high 70s to upper 80s for better systems. Participants rated the spatialized network display with an average score of 78 and the self-organizing map interface component with 81.

DISCUSSION

We employed the HDS as an example of a semi-structured text archive in the humanities for our case study. We implemented two spatialized display types to explore how spatio-temporal and thematic information extracted from unstructured and semi-structured text archives in the humanities can be efficiently and effectively transformed and visualized to facilitate the uncovering and exploration of spatio-temporal and thematic structures and interconnections. Our proposed use of graphs (i.e., networks) and (self-organizing) maps, recommended by Moretti (2005) for *distant reading* in the humanities, indeed supported historians in gaining new insights from uncovered latent spatio-temporal and thematic structures and relationships in the HDS text archive. In particular, we found that participants could meaningfully interpret the distance-similarity metaphor in both the spatialized network and thematic SOM displays, allowing exploration of spatio-temporal and thematic interconnections in the HDS articles, and matching this overview information with their own readings

of HDS articles. Our work extends previous empirical spatialization research on text document repositories that quantitatively assessed the perceptual properties of static spatialized views. Specifically, Montello et al. (2003), Fabrikant et al. (2006) and (2014) evaluated how perceptual properties of the display modulate the intuitive understanding of the distance-similarity metaphor, visualized in static 2D and 3D spatialized views of text archives. They found that humans interpret the spatial separation of items in a display to be a metaphor for dissimilarity in static spatialized views. However, how the semantics of the shown data items (i.e., the content of text documents) is understood in spatialized views was not evaluated. In this case study, we empirically evaluated interactive spatialized views that follow the tested visualization principles of the distance-similarity metaphor. We additionally gave target users the opportunity to not only explore spatio-temporal and thematic structures and interconnections, revealed from a document collection about Swiss history (i.e., their semantics), but also offer a close reading of the individual data items. We illustrated that the distance-similarity metaphor in spatializations operates with a *real-world* application for our target users in the humanities.

While spatialized network displays have also been successfully implemented for example to visualize social relationships extracted from text documents in the digital humanities (e.g., Bingenheimer et al., 2011), this community still lacks empirically validated proof-of-concept spatializations with target users of spatio-temporal information, automatically extracted from large semi-structured text collections. Our evaluation results illustrate that network spatializations and SOMs support target users in the humanities in their information-seeking process beyond close reading of individual text documents, in identifying and analyzing latent spatio-temporal and thematic structures and interconnections in large digital text archives. Our case study highlights the potential benefits of including spatialized displays as complementary visual aids for distant reading, to explore space, time, and theme in the humanities.

Our empirical results are, however, limited to the assessment of two specific, albeit common visualization techniques. We thus did not assess whether the selection of networks and SOMs themselves might have had a potential influence on the information-seeking process of target users in the humanities. We did not evaluate whether the use of other visualization techniques could influence target users' sense making of the document collection. For example, tree visualizations, as proposed by Moretti (2005), also emphasize hierarchical structures, and could be an alternative to spatialized networks. However, tree maps are not as powerful in highlighting relationships between single data items (allowing close reading) as spatialized networks (Shneiderman, 1996). It would be interesting to assess whether and how different visualization methods and varying graphic characteristics of visualizations might influence the type and quality of insights

target users might gain from a document collection. We also did not evaluate whether the interactive spatialized displays might lead to different insights, compared to simply reading individual HDS article texts only.

Our second research question related to whether and how information about space, time, and theme buried in unstructured and semi-structured text archives could be made available to information seekers in the humanities to support sense-making, and the generation of new insights about these text archives. We detailed an iterative, user-centered interface design and evaluation approach to answer this question. As recommended by Lewis and Rieman (1993) and Roth et al. (2015), the early involvement of target users in the evaluation of a graphical user interface is important for success. The high system usability scores (SUS) we obtained with our proof-of-concept is a very encouraging result in this direction. To increase the potential utility of using spatialized views for document exploration, we followed suggestions offered by Moretti (2005) and Jockers (2013) coupling *distant reading* and *close reading* views in our interactive user interface to support the exploratory information-seeking process for humanities scholars. The design of the user interface was informed by Shneiderman's (1996) information-seeking mantra (i.e., overview first, zoom and filter, then details-on-demand). Indeed, participants seem to closely follow Shneiderman's (1996) information-seeking mantra both, when using the *distant* and *close reading* views offered for this reason in the interface to gain new insights about space, time, and thematic content of the document collection. Interactive information visualizations that already contain combined *distant* and *close reading* views to explore text archives in the humanities, and that are also designed with the information-seeking mantra in mind, can be found in digital humanities literature (e.g., Hinrichs et al., 2015). We extend prior research to include user-tested spatialized views in an interactive web interface, coupling *distant* and *close reading* to explore spatio-temporal and thematic information retrieved from a semi-structured text archive of Swiss history.

Our project not only extends similar research in the digital humanities, but also contributes to ongoing work in geographic information search and retrieval. We provide empirical evidence for Ballatore et al.'s (2015) contention that spatial, temporal, and thematic information always interact in search, and thus that these three dimensions need to be integrated in information retrieval and search systems. Furthermore, we respond to calls in geographic information science and spatial information retrieval for the systematic evaluation of exploratory search in the context of unstructured and semi-structured data archives (Ballatore et al., 2015). We successfully demonstrate that an iterative and systematic user-centered design and evaluation framework is key to building useable and useful web-based spatialized search tools.

An important limitation in our case study is related to a core assumption made developing the spatialized networks. Based on prior work by Hecht and

Raubal (2008) and Salvini and Fabrikant (2016), we assume that a relationship between two toponyms exists, if these two toponyms co-occur often in the same HDS articles. We treat articles as *bags of words,* thus disregarding syntax (i.e., word order, sentence structure, etc.) in the analysis. To consider syntax in future studies textual proximity could be used to weigh the strength of word relationships: the more often toponyms co-occur in the same sentences (or paragraphs, etc.) in an article, the stronger the relationship between the respective toponyms.

CONCLUSIONS AND OUTLOOK

With this theoretically-driven, empirical research project we are able to illustrate how spatio-temporal and thematic information in large online digital text archives in the humanities can be retrieved, transformed, and visualized in a web interface such that latent spatio-temporal and thematic structures and interconnections can be interactively explored by target users in the humanities. We chose a semi-structured online digital text archive in the humanities as a case study and presented a comprehensive approach to retrieving information about space, time, and theme automatically from the text documents using established (geographic) information retrieval methods. We then transformed and reorganized the retrieved information and visualized latent spatio-temporal and thematic patterns and relationships in spatialized displays that we empirically tested with users. To do so, we applied a user-centered design and evaluation strategy iteratively designing, developing, and assessing spatialized web interface components with users to access a semi-structured document collection of Swiss history. We could show that target users in the humanities can generate new insights about spatio-temporal and thematic structures and interconnections in the digital text archive. Our translational approach introduces established methods in GIScience to expand (digital) humanities research with advanced computing technologies.

Our approach could be applied to other language versions (i.e., French, Italian, Rhaeto-Romanic) of the HDS, to assess the stability and applicability of the methods and techniques in a multilingual context. Of course, to explore how the applied methods scale and transfer to other data domains and data sources in the humanities or beyond, they need to be analyzed in a comparative fashion.

In closing, we hope to have provided a direct answer on how interesting, meaningful, and relevant information buried in large online text archives can be accessed and explored. This is relevant to all of us in the information society as the amount of data stored in large online archives continues to increase exponentially.

AUTHOR NOTE

Correspondence concerning this article should be addressed to Sara I. Fabrikant, Department of Geography, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland.

ACKNOWLEDGEMENTS

We would like to thank Curdin Derungs, Damien Palacio, Jannik Strötgen, and Julian Zell, who specifically helped us to implement the GIR portions of this research. Funding by the Canton of Zurich is gratefully acknowledged.

REFERENCES

Adams, Benjamin and Gahegan, Mark (2016), 'Exploratory Chronotopic Data Analysis', in Jennifer A. Miller, David O'Sullivan, and Nancy Wiegand (eds.), *Geographic Information Science: 9th International Conference, GIScience 2016, Montreal, QC, Canada, September 27–30, 2016, Proceedings* (Cham, Switzerland: Springer International Publishing), 243–58.

Andrienko, G., et al. (2007), 'Geovisual analytics for spatial decision support: Setting the research agenda', *International Journal of Geographical Information Science*, 21 (8), 839–57.

Ballatore, Andrea, et al. (2015), 'Spatial Search, Final Report'. < http://www.escholarship.org/uc/item/33t8h2nw > .

Bangor, Aaron, Kortum, Philip T., and Miller, James T. (2008), 'An Empirical Evaluation of the System Usability Scale', *International Journal of Human–Computer Interaction*, 24 (6), 574–94.

Bingenheimer, Marcus, Hung, Jen-Jou, and Wiles, Simon (2011), 'Social network visualization from TEI data', *Literary and Linguistic Computing*, 26 (3), 271–78.

Blondel, Vincent D., et al. (2008), 'Fast unfolding of communities in large networks', *Journal of Statistical Mechanics: Theory and Experiment*, 2008 (10), P10008.

Brooke, John (1996), 'SUS – A 'quick and dirty' usability scale', in Patrick W. Jordan, et al. (eds.), *Usability evaluation in industry* (London, UK: Taylor & Francis), 189–94.

Bruggmann, André (2017), 'Visualization and Interactive Exploration of Spatio-Temporal and Thematic Information in Digital Text Archives', (doctoral dissertation). University of Zurich.

Bruggmann, André and Fabrikant, Sara I (2016), 'How does GIScience support spatio-temporal information search in the humanities?', *Spatial Cognition & Computation*, 16 (4), 255–71.

Dearholt, Donald W. and Schvaneveldt, Roger W. (1990), 'Properties of pathfinder networks', in Roger W. Schvaneveldt (ed.), *Pathfinder associative networks* (Norwood, NJ, USA: Ablex Publishing Corp.), 1–30.

Derungs, Curdin and Purves, Ross S. (2014), 'From text to landscape: locating, identifying and mapping the use of landscape features in a Swiss Alpine corpus', *International Journal of Geographical Information Science*, 28 (6), 1272–93.

Esri (2017), 'ArcGIS Online', (http://www.esri.com/software/arcgis/arcgisonline).

Fabrikant, Sara I and Skupin, André (2005), 'Cognitively Plausible Information Visualization', in Jason Dykes, Alan M. MacEachren, and Menno-Jan Kraak (eds.), *Exploring Geovisualization* (Amsterdam, Netherlands: Elsevier), 667–90.

Fabrikant, Sara I, Montello, Daniel R, and Mark, David M (2006), 'The distance similarity metaphor in region-display spatializations', *IEEE Computer Graphics and Applications*, 26 (4), 34–44.

Fabrikant, Sara I, Montello, Daniel R, and Mark, David M (2010), 'The natural landscape metaphor in information visualization: The role of commonsense geomorphology', *Journal of the American Society for Information Science and Technology*, 61 (2), 253–70.

Fabrikant, S. I., Maggi, S., Montello, D. R. (2014). 3D network spatialization: Does it add depth to 2D representations of semantic proximity? Proceedings, GIScience 2014, Wien, Sep. 23–26, 2014, Duckham, M., Pebesma, E., Stewart, K. (eds.), Lecture Notes in Computer Science (LNCS) 8728 Springer, Berlin, Germany: 34–47.

HDS 'Historical Dictionary of Switzerland', < http://www.hls-dhs-dss.ch > , last accessed, 30 Sept, 2019.

Hecht, Brent and Raubal, Martin (2008), 'GeoSR: Geographically Explore Semantic Relations in World Knowledge', in Lars Bernard, Anders Friis-Christensen, and Hardy Pundt (eds.), *The European Information Society: Taking Geoinformation Science One Step Further* (Berlin, Heidelberg, Germany: Springer Berlin Heidelberg), 95–113.

Hinrichs, Uta, et al. (2015), 'Trading Consequences: A Case Study of Combining Text Mining and Visualization to Facilitate Document Exploration', *Digital Scholarship in the Humanities*, 30 (suppl 1), i50–i75.

Jockers, Matthew L. (2013), *Macroanalysis: Digital Methods & Literary History*, eds Susan Schreibman and Raymond C. Siemens (Urbana, IL, USA: University of Illinois Press).

Jones, Christopher B. and Purves, Ross S. (2008), 'Geographical information retrieval', *International Journal of Geographical Information Science*, 22 (3), 219–28.

Jones, Rosie, et al. (2008), 'Geographic intention and modification in web search', *International Journal of Geographical Information Science*, 22 (3), 229–46.

Kaplan, Frédéric (2015), 'A Map for Big Data Research in Digital Humanities', *Frontiers in Digital Humanities*, 2.

Lewis, Clayton and Rieman, John (1993), *Task-centered user interface design: A practical introduction* (Boulder, CO, USA: University of Colorado, Boulder).

Montello, Daniel R., et al. (2003), 'Testing the First Law of Cognitive Geography on Point-Display Spatializations', in Walter Kuhn, Michael F. Worboys, and Sabine Timpf (eds.), *Spatial Information Theory. Foundations of Geographic Information Science: International Conference, COSIT 2003, Kartause Ittingen, Switzerland, September 24–28, 2003. Proceedings* (Berlin, Heidelberg: Springer Berlin Heidelberg), 316–31.

Moretti, Franco (2005), *Graphs, Maps, Trees: Abstract Models for Literary History* (London, UK: Verso).

North, Chris (2006), 'Toward measuring visualization insight', *IEEE Computer Graphics and Applications*, 26 (3), 6–9.

Purves, Ross S., et al. (2018), Geographic information retrieval: Progress and challenges in spatial search of text. Foundations and Trends in Information Retrieval 12 (2–3), 164–318.

Roth, Robert E., Ross, Kevin S., and MacEachren, Alan M. (2015), 'User-Centered Design for Interactive Maps: A Case Study in Crime Analysis', *ISPRS International Journal of Geo-Information*, 4, 262–301.

Robinson, A. (2017). Geovisual Analytics. *The Geographic Information Science & Technology Body of Knowledge* (3rd Quarter 2017 Edition), John P. Wilson (ed.). DOI: 10.22224/gistbok/2017.3.6.

Salvini, Marco M. and Fabrikant, Sara I. (2016), 'Spatialization of user-generated content to uncover the multirelational world city network', *Environment and Planning B: Planning and Design*, 43 (1), 228–48.

Shneiderman, Ben (1996), 'The eyes have it: a task by data type taxonomy for information visualizations', *1996 IEEE Symposium on Visual Languages* (Boulder, CO, USA), 336–43.

Skupin, André and Agarwal, Pragya (2008), 'Introduction: What is a Self-Organizing Map?', in Pragya Agarwal and André Skupin (eds.), *Self-Organising Maps: Applications in Geographic Information Science* (Chichester, UK: John Wiley & Sons), 1–20.

Steyvers, Mark and Griffiths, Tom (2007), 'Probabilistic Topic Models', in Thomas K Landauer, et al. (eds.), *Handbook of Latent Semantic Analysis* (Mahwah, NJ, USA: Lawrence Erlbaum Associates), 427–48.

Strötgen, Jannik and Gertz, Michael (2013), 'Multilingual and cross-domain temporal tagging', *Language Resources and Evaluation*, 47 (2), 269–98.