GlamMap: Geovisualization for e-Humanities

T. Castermans, B. Speckmann, K. Verbeek, M. A. Westenberg, A. Betti and H. van den Berg

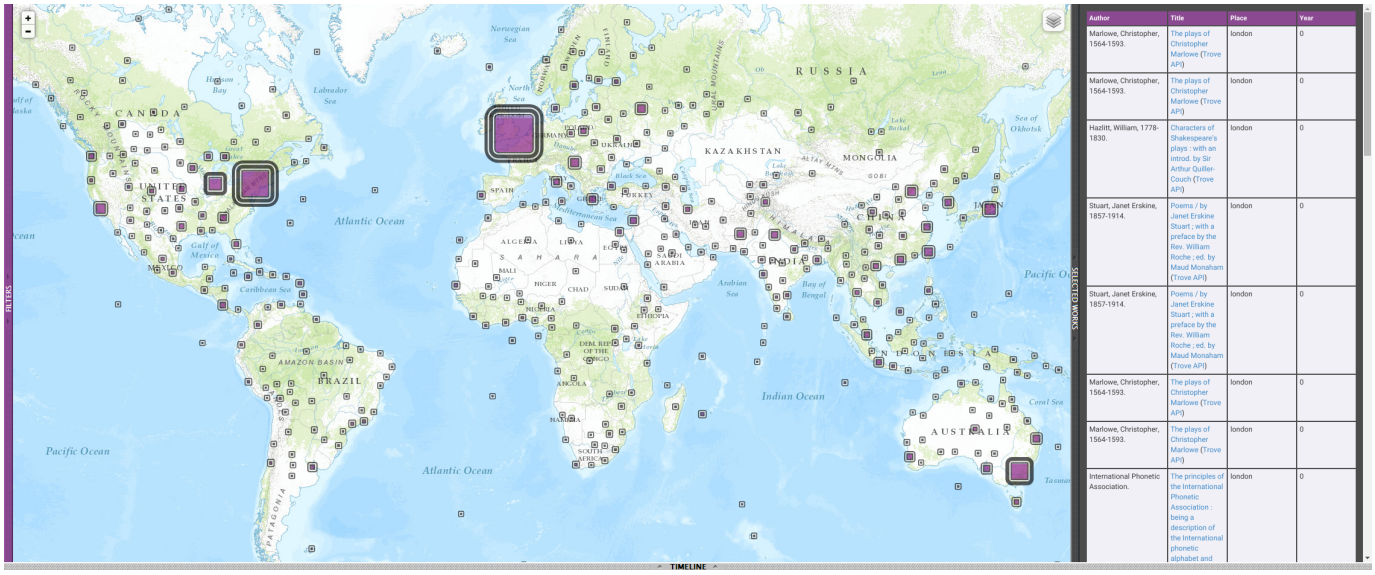


Fig. 1. Visualization of the Trove dataset [1] consisting of about 60 million records in GlamMap.

Abstract—This paper presents GlamMap, a visualization tool for large, multi-variate georeferenced humanities data sets. Our approach visualizes the data as glyphs on a zoomable geographic map, and performs clustering and data aggregation at each zoom level to avoid clutter and to prevent overlap of symbols. GlamMap was developed for the Galleries, Libraries, Archives, and Museums (GLAM) domain in cooperation with researchers in philosophy. We demonstrate the usefulness of our approach by a case study on history of logic, which involves navigation and exploration of 7100 bibliographic records, and scalability on a data set of sixty million book records.

Index Terms—Clustering, image generation, information retrieval

1 Introduction

GlamMap is a visualization tool for georeferenced data sets in the humanities field. While visualization has become common practice in many sciences, researchers in humanities have limited access to specialized tools relevant to their research questions. Also, end users in this domain, more generally referred to as Galleries, Libraries, Archives, and Museums (GLAM), are still mainly working with text-based interfaces to access and maintain their collections.

It was shown that visualizing GLAM data in its geographic context has many benefits to the digital humanities community [4]. In this paper, we extend upon that work, and address scalability. Our approach visualizes the data as glyphs on a zoomable geographic map. The data is aggregated algorithmically to prevent overlap of glyphs. We have designed a scalable system, so that GlamMap can work with very large databases, as demonstrated in Section 5.2.

- *T. Castermans, B. Speckmann, K. Verbeek and M. A. Westenberg are with TU Eindhoven, Netherlands. E-mail: {t.h.a.castermans, b.speckmann, k.a.b.verbeek, m.a.westenberg}@tue.nl.*
- *A. Betti and H. van den Berg are with University of Amsterdam, Netherlands. E-mail: {a.betti, h.vandenberg}@uva.nl.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

2 Related Work

The ability to efficiently analyze and explore a large number of documents in a digital library is an important objective in many domains[1]. Standard textual interfaces are not sufficient for this purpose and suitable visual interfaces are needed [5, 6]. As a result, many visualization tools have been developed to visualize digital libraries, in many different ways. Most of these approaches rely on a hierarchy in the metadata, such as the Dewey Decimal classification or other schemes. Some tools produce plots on meaningful axes to visualize the data [16, 20]. Compact visualization of hierarchies can also be attained by treemaps, which are confined to a predefined rectangular region [10]. Furthermore, some systems use relations between documents to visualize the data as a network of entities [23]. Another area that has received much attention is topic visualization over large document collections (see e.g. [11, 12, 17] for recent overviews). Visual topic analysis systems help users to explore and understand topic evolutions. Sets of independent topics can be visualized by tag clouds [11], but this technique is less suitable for tracking evolution of multiple, dependent topics. Cui *et al.* [12] have proposed a river-flow-based visual metaphor to handle the latter case. The Bohemian Bookshelf [24] is a system that is aimed at discovery in digital book collections. The system was designed for use in public information displays, so an open question is whether the approach is usable in a scientific context, and also whether it scales to very large document collections.

The types of visualizations used by these tools often depend strongly on the target domain. Koch *et al.* [15] present a system for searching

[1] In this section and in Section 5.1, we reuse work and lift sentences from [4].

and analyzing patents, which focuses on query refinement. Brehmer *et al.* [7] present an application aimed at investigative journalists, which focuses on exploratory content-based analysis of large document collections. In some domains, as is the case for our target domain, geographic metadata plays an important role and users must be able to specify geographical scope. Map-based visualizations can be very powerful tools to understand large amounts of georeferenced data. However, very few map-based visualizations exist for bibliographic data. GeoVIBE [9] is a tool that uses a map-based visualization to show geographical information linked to a collection of documents. However, its functionalities are not scalable and do not match the needs of our users. Tools created by DPLA[2], OCLC[3] and Europeana[4] similary have proven to be useful to users, but are very limited in terms of visualization techniques.

In GlamMap we show items at their respective locations. Since every document is represented by a symbol, our problem resembles that of dynamic map labeling [2]. However, instead of omitting items that cannot be shown due to overlap, we aggregate nearby items into disjoint glyphs [27]. Note that this is different from grid-based aggregation[5]. Similar techniques have been used before (see e.g. [19], which includes a good overview), and related approaches have also been applied to network exploration (see e.g. [26]).

3 Problem Analysis

The input of our application consists of a set of records describing data from the humanities domain. Typically, such data sets consists of bibliographical records, collections of artifacts, or other, more generic, multivariate data. In this paper, we focus on data that is georeferenced. To give an idea how such data sets are organized, we describe the FRBR (Functional Requirements for Bibliographic Records [14]) standard. It organizes the data hierarchically using the following four levels.

1. Works (e.g. *The Odyssey*)
2. Expressions (e.g. Fagles' translation of 1.)
3. Manifestations (e.g. a particular paperback edition of 2.)
4. Items (e.g. a copy of 3.)

The target users mainly consist of two groups: (1) GLAM data providers, in particular librarians and library data providers, and (2) researchers in the humanities.

Librarians and library data providers. An important task for librarians is *collection assessment*. We can distinguish between internal and external collection assessment. For internal collection assessment, a librarian may want to figure out if a particular topic is duly covered in his/her library. To reduce costs, the library cooperates with other libraries in the neighborhood, so that this library itself does not need all relevant works for a particular topic, as long as neighboring libraries can fill the gap. The librarian would therefore like to quickly determine how well a topic is covered in the neighboring libraries.

For external collection assessment, consider a library user that wants to find a particular work, possibly with certain special characteristics. This is typically the kind of query a library data provider aims to support. Naturally, it is relevant to the user if this work is available at one of the nearby libraries, or to find the closest library that holds this work. Furthermore, the user should be able to quickly see if this work satisfies the required special characteristics.

Researchers. Researchers in the humanities would like to explore the ever-growing quantity of data nowadays available to make interesting discoveries. For example, researchers interested in the history of a certain field may want to study the dissemination of scientific knowledge within that field (see e.g. [25]). Unfortunately, in the way that this data is commonly available (e.g. WorldCat), it is impossible to explore substantial amounts of book records easily and quickly for research purposes. Thus, researchers in the humanities require tools that aid them in the first phase of their research, i.e. hypotheses forming (literature search, corpus building and initial content exploration). A visual representation of the data with interactive exploration may help the researcher to discover patterns in the data that can lead to new hypotheses. Additionally, the application can further assist the researcher in verifying these hypotheses.

3.1 Tasks

The high-level objectives described in the previous section, can be decomposed into a number of generic tasks. These are described in the main objectives of librarians and researchers of bibliographic data (but generalize to other scenarios).

T1-Selecting a relevant subset. The user must be able to restrict or filter the dataset based on several types of metadata so that only the relevant parts of the data are shown. Librarians may only be interested in a particular topic. Researchers may only want to investigate a certain subset of authors, or only consider a particular time period.

T2-Investigating individual items. It must be possible to not only find, but also access all metadata related to a single item in the dataset. Additionally, the hierarchical relation between the different levels of the FRBR standard should be clear to the user. For example, it must be possible to find all copies (items) of a particular work.

T3-Browsing all items in a geographic region. The user must be able to browse all items that are available at a particular location. The application should also offer a good overview of the items available in a particular geographic region. Note that the meaning of the location may depend on the data (physical location, location of publisher, etc.).

T4-Analyzing the distribution of metadata. Given a particular type of metadata, the user should be able to get a quick overview of how this variable is distributed over the different items in the dataset.

Note that tasks T1, T2, and T3 are useful to both groups of users. Task T4 is mostly aimed at researchers, and could prove to be a crucial tool in discovering interesting patterns.

4 Approach

We provide a map-based visualization of the data, for which we rely on existing technology such as ArcGIS and Leaflet. GlamMap visualizes the data as glyphs in a layer that is superimposed on the map.

Constraints. In principle, we display a separate glyph for every item in the data in order to facilitate task T2. Aggregation of the data is allowed, as long as it is possible to retrieve a single item through interaction. Because our application must be able to support datasets of varying sizes, ranging from very small datasets with only 10-20 records to very large datasets with tens or even hundreds of millions of records (like Trove [1, 3] and WorldCat), we aggregate data items if their glyphs overlap in the visualization. This simplifies performing task T3 and T4. To effectively support these tasks for large datasets, we need to ensure that any basic type of interaction like panning and zooming works at interactive speeds, regardless of the size of the dataset. As a result, any type of data aggregation we use while zooming must be very efficient. Additionally, the aggregation must be consistent so that the mental map of the user is preserved while interacting with the visualization. To aid the user in this, we provide a preview of the change in clustering at the next zoom level if the user mouses over a glyph.

A further consideration is the size of the symbols. Symbols must be large enough to be readable, but not so large that all symbols must be aggregated into one symbol. This essentially boils down to achieving the right "fullness" of the map [21, p. 315]. Therefore, symbol sizes must be chosen carefully to obtain a satisfactory visualization.

The symbols must also be visually separable from the underlying map, which means that there should be a large enough contrast break between the symbol and the map.

[2]DPLA by county and state, `http://dp.la/apps/14/`.

[3]WorldCat, `http://www.worldcat.org/`.

[4]Europeana Foundation, `http://www.europeana.eu/portal/`.

[5]Grid-based aggregation is supported in ElasticSearch, `https://www.elastic.co/guide/en/elasticsearch/reference/current/search-aggregations-bucket-geohashgrid-aggregation.html`. It is also used in Nanocubes, `http://www.nanocubes.net/`.

(a) (b)

Fig. 2. (a) A glyph showing one square for each of eighteen items in its interior. One item (square) is colored blue and the others purple. (b) A glyph with its items categorized sequentially into five categories.

Clustering. Grouping of items is necessary if they map to the same location, or because the corresponding glyphs would overlap in screen space. The latter situation is detected and resolved algorithmically.

As mentioned earlier, we aggregate glyphs at different locations in order to avoid overlap. For a particular zoom level, this can be seen as a clustering of the glyphs, where two glyphs should be in the same cluster if they would otherwise overlap. This clustering must be both consistent and efficient. If the user pans, then the clustering should not change. Furthermore, if two glyphs are in the same cluster at a particular zoom level, then they should remain in the same cluster when zooming in further. To enforce this, we globally compute a hierarchical clustering for all glyphs on all zoom levels.

To compute this hierarchical clustering, we use an agglomerative approach. That is, we start with individual glyphs and merge glyphs when they start to overlap as we zoom out. However, at a particular zoom level, we can have multiple overlaps among glyphs. We could simply put every connected set of glyphs into one cluster, but this would be wasteful: merging two glyphs may actually remove overlaps. Therefore, at every zoom level, we merge overlapping glyphs incrementally. We first merge the two glyphs with the largest area of overlap (as in [19]) and repeat this procedure until all glyphs are disjoint. These glyphs are then used to compute the clustering for the next zoom level.

The clustering hierarchy allows efficient updates on the clustering while panning and zooming, but it must be computed upon loading the map. Furthermore, it must be recomputed whenever the data is filtered. It is therefore important that the hierarchical clustering can be computed efficiently. A naive approach would lead to a running time of $O(n^3)$ for n locations: there can be at most $n-1$ merges in total and computing the next pair of glyphs to be merged takes $O(n^2)$ time by trying all pairs. This running time is unacceptable for our use cases, which can contain thousands or tens of thousands of locations. To remedy this, we have devised an algorithm that computes the hierarchical clustering in $O(n^2)$ time. See [8] for details.

Glyph design. The shape of a glyph must be simple to easily understand the density of items on the map. We use rounded rectangles as they are visually pleasing, and enable users to estimate the area. An alternative option were circles, because they are compact and visually stable [13] and preferred by users [22]. However, estimating areas of circles is perceptually hard.

The glyph interior consists of a heat map, in which each square corresponds to an item. The color represents an item attribute for categorical or numerical attributes. If the number of items is small, the individual squares become visible (see Fig. 2(a)). Items are sorted by category so it is easy to see the distribution of items over categories. This approach can be scaled to large numbers of items, as it allows to compare the relative contributions of each of the categories (as in Fig. 2(b) for example). The glyph is somewhat translucent so that map details are not completely obscured.

To contrast the glyph from the map, a gray border is drawn. This border may be used to encode additional information. An example of this is the compression level. For large datasets, such as Trove (see Section 5.2), linearly scaling the size of the glyphs according to the number of items they contain is not viable. Glyphs that represent many items would simply become too large and cover most of the map. To deal with this issue, we introduce compression levels for the glyphs. The compression level of a glyph is determined by the number of items it represents; two thresholds determine the three levels. Compression level 0, 1, and 2 scale the glyph interior with factors 1, $5/6$ and $2/3$, respectively. These values have been determined experimentally. The compression level is encoded in the glyph by adding additional border layers. This gives the impression that the glyph interior is pushed inwards, and therefore is shown smaller than it actually is. Fig. 1 shows glyphs of all three levels: the glyphs corresponding to London and Washington are drawn at compression level 2, several glyphs are drawn at compression level 1 in the US, Australia, and Europe, and most glyphs are drawn at their actual, uncompressed size.

5 Results

The utility of geovisualization for librarians and historians of logic has been demonstrated recently [4]. In the following, we will repeat a use case from this paper. We will also demonstrate that GlamMap scales to huge datasets containing millions of items. Both described results can be found online: the Risse data set (`http://glammap.win.tue.nl/glamdev/maps/1`) and Trove data set (`http://glammap.win.tue.nl/glamdev/maps/5`).

5.1 Use for historians of logic

GlamMap has visualized 7,100 bibliographic records from books in logic published in Europe between 1700 en 1940 (Fig. 1 in [4]). This dataset was obtained from Wilhelm Risse's Bibliographica Logica [18]. This visualization provides historians of logic with a quick overview of which books on logic were published when and where. By inspecting the visualization, the historians learned that famous works on logic, such as, for example, Bernard Bolzano's Wissenschaftslehre, were published relatively few times (according to Risse, the Wissenschaftslehre was published 3 times from 1700 to 1940). By contrast, textbooks on logic written by the little known English theologian Isaac Watts, who is little studied by historians of logic, were quite often published in the 18th and 19th century (according to Risse, 39 books by Watts were published between 1725 and 1875). This result suggests that, as is the case in the twenty-first century, important ideas in 18th and 19th century logic were often communicated through popular and simplified textbooks, see Fig. 3. Such a result is of great importance for historians of logic. It shows them which resources to study in order to understand the communication and dissemination of ideas [4].

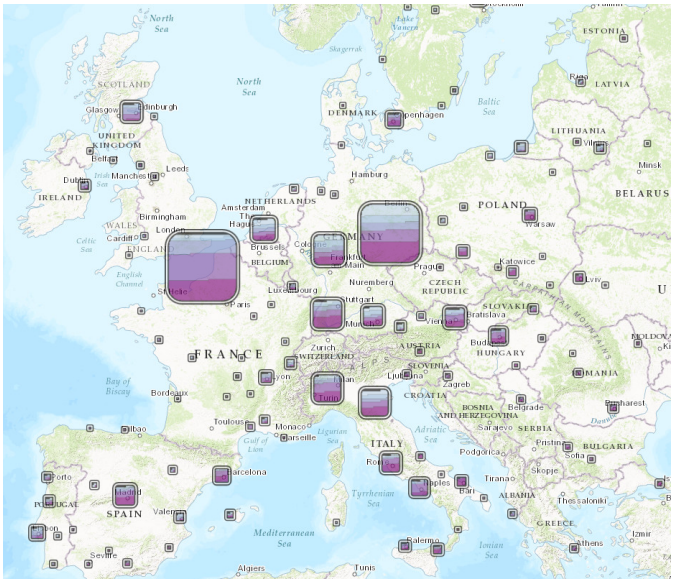
5.2 Scalability

Visualization of large bibliographic databases is very helpful for historians to identify both known and unknown texts. The larger the database, the better; ideally, historians want to have access to all bibliographic data of writings published in a certain period. GlamMap can work with very large databases, such as Trove [1, 3], which contains about 60 million bibliographic records (shown in Fig. 1). This allows historians to gain a quick overview of geographic distribution of works, total number of publications in certain historical periods, and insight into publications (in original language and translations) by a specific author, for example, to identify novel research data.

6 Conclusion

We have presented GlamMap, a visualization tool for large georeferenced humanities data sets. Our approach visualizes the data as glyphs on a zoomable geographic map, and performs clustering and data aggregation at each zoom level to avoid clutter and to prevent overlap of symbols. The use cases demonstrate that GlamMap allows domain experts to explore the ever-growing quantity of data, and to discover interesting patterns and form new hypotheses inspiring further research.

We have demonstrated GlamMap at various workshops and venues involving humanities researchers and users of GLAM data. Researchers judge the tool to be attractive, because it is considered more flexible, insightful, beautiful, and faster than other available approaches. Librarians see use for GlamMap in portals provided by aggregators and consortia such as Europeana, who can contribute to spreading awareness among users as to the existence of more advanced data visualization.



(a)

(b)

Fig. 3. (a) Textbooks in the Risse dataset. (b) A timeline shows the amount of items per year, works as a color legend, and allows the user to specify a range filter. Reproduced from [4].

Although GlamMap is able to visualize large data sets interactively, there are still some limitations and aspects that need further improvement. Scalability is a concern, both visually and algorithmically. The clustering approach is efficient, but not fast enough to deal with huge data sets at interactive rates. As filtering can remove items from or add items to the current view, this needs to be improved by allowing dynamic updates to the clustering. This is a challenging algorithmic problem, however. So far, we are not aware of any research that addresses this aspect. As part of future work, we plan to improve the scalability of GlamMap. We collaborate with OCLC to develop a GlamMap-based interface for WorldCat, which holds hundreds of millions of bibliographic records.

Finally, it is interesting to extend GlamMap to handle more complex geographic relations. In particular, we like to consider items that relate to multiple locations or even items that move. For example, consider a painting that is moved between several museums. New visualization techniques will be required to facilitate the discovery of interesting patterns in such rich datasets that may be of great interest to researchers in the humanities.

Acknowledgments

The Netherlands Organisation for Scientific Research (NWO) is supporting B. Speckmann under project no. 639.023.208, K. Verbeek under project no. 639.021.541, and A. Betti, H. van den Berg, and T. Castermans under project no. 314.99.117.

References

[1] Trove. `http://trove.nla.gov.au/`.

[2] K. Been, E. Daiches, and C. Yap. Dynamic map labeling. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):773–780, 2006.

[3] A. Betti, T. Castermans, B. Speckmann, K. Verbeek, and H. van den Berg. GlamMapping Trove. In *Proceedings of the VALA 18th Biennial Conference and Exhibition*. Melbourne, 2016.

[4] A. Betti, D. H. P. Gerrits, B. Speckmann, and H. van den Berg. GlamMap: Visualising library metadata. In *Proceedings of the VALA 17th Biennial Conference and Exhibition*. Melbourne, 2014.

[5] K. Börner. iScape: A collaborative memory palace for digital library search results. In *Proceedings of the 9th International Conference on Human-Computer Interaction*, pp. 1160–1164, 2001.

[6] K. Börner and C. Chen. Visual interfaces to digital libraries: Motivation, utilization, and socio-technical challenges. In K. Börner and C. Chen, eds., *Visual Interfaces to Digital Libraries*, vol. 2539 of *Lecture Notes in Computer Science*, pp. 1–9. Springer Berlin Heidelberg, 2002.

[7] M. Brehmer, S. Ingram, J. Stray, and T. Munzner. Overview: The design, adoption, and analysis of a visual document mining tool for investigative journalists. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2271–2280, 2014.

[8] J. Brekelmans. Interactive geographic visualization of very large GLAM data. Master's thesis, Technische Universiteit Eindhoven, 2015.

[9] G. Cai. Geovibe: A visual interface for geographic digital libraries. In K. Börner and C. Chen, eds., *Visual Interfaces to Digital Libraries*, vol. 2539 of *Lecture Notes in Computer Science*, pp. 171–187. Springer Berlin Heidelberg, 2002.

[10] E. Clarkson, K. Desai, and J. Foley. Resultmaps: Visualization for search interfaces. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1057–1064, 2009.

[11] C. Collins, F. Viegas, and M. Wattenberg. Parallel tag clouds to explore and analyze faceted text corpora. In *Proceedings of the 4th IEEE Symposium on Visual Analytics Science and Technology (VAST '09)*, pp. 91–98, 2009.

[12] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong. Textflow: Towards better understanding of evolving topics in text. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2412–2421, 2011.

[13] B. D. Dent. *Cartography: Thematic Map Design*. McGraw-Hill, 6th ed., 2008.

[14] IFLA. Functional requirements for bibliographic records: final report. Technical report, IFLA Study Group on the Functional Requirements for Bibliographic Records, München: KG Saur, 1998.

[15] S. Koch, H. Bosch, M. Giereth, and T. Ertl. Iterative integration of visual insights during scalable patent search and analysis. *IEEE Transactions on Visualization and Computer Graphics*, 17(5):557–569, 2011.

[16] L. Marks, J. A. Hussell, T. M. McMahon, and R. E. Luce. Activegraph: A digital library visualization tool. *International Journal on Digital Libraries*, 5(1):57–69, 2005.

[17] D. Oelke, H. Strobelt, C. Rohrdantz, I. Gurevych, and O. Deussen. Comparative exploration of document collections: a visual analytics approach. *Computer Graphics Forum*, 33(3):201–210, 2014.

[18] W. Risse. *Bibliographia logica*. Georg Olms, 1965.

[19] R. Scheepens, H. van de Wetering, and J. J. van Wijk. Non-overlapping aggregated multivariate glyphs for moving objects. In *Proceedings of the 7th IEEE Pacific Visualization Symposium*, pp. 17–24, 2014.

[20] B. Shneiderman, D. Feldman, A. Rose, and X. F. Grau. Visualizing digital library search results with categorical and hierarchical axes. In *Proceedings of the 5th ACM Conference on Digital Libraries (DL '00)*, pp. 57–66. ACM, 2000.

[21] T. A. Slocum, R. B. McMaster, F. C. Kessler, and H. H. Howard. *Thematic Cartography and Geovisualization*. Prentice Hall, 3rd ed., 2008.

[22] T. A. Slocum, J. Robert S. Sluter, E. C. Kessler, and S. C. Yoder. A qualitative evaluation of MapTime, a program for exploring spatiotemporal point data. *Cartographica*, 39(3):43–68, 2004.

[23] J. Stasko, C. Gorg, Z. Liu, and K. Singhal. Jigsaw: Supporting investigative analysis through interactive visualization. In *Proceedings of the 2nd IEEE Symposium on Visual Analytics Science and Technology (VAST '07)*, pp. 131–138, 2007.

[24] A. Thudt, U. Hinrichs, and S. Carpendale. The bohemian bookshelf: Supporting serendipitous book discoveries through information visualization. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*, pp. 1461–1470. ACM, 2012.

[25] H. van den Berg, G. Parra, A. Jentzsch, A. Drakos, and E. Duval. Studying the history of philosophical ideas: Supporting research discovery, navigation, and awareness. In *Proceedings of the 14th International Conference on Knowledge Technologies and Data-driven Business (i-KNOW '14)*, pp. 12:1–12:8. ACM, 2014.

[26] C. Vehlow, T. Reinhardt, and D. Weiskopf. Visualizing fuzzy overlapping communities in networks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2486–2495, 2013.

[27] M. O. Ward. Multivariate data glyphs: Principles and practice. In *Handbook of Data Visualization*, Springer Handbooks of Computational Statistics, pp. 179–198. Springer Berlin Heidelberg, 2008.