GlamMap: Geovisualization for E-Humanities

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

Publication date
2016

Document Version
Final published version

Published in
1st Workshop on Visualization for the Digital Humanities

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
GlamMap: Geovisualization for e-Humanities

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

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.

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. 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
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 DPLA2, OCLC3 and Europeana4 simiary 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 aggregation5. 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 consist 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.
**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 $O(n^2)$ 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 $O(n^2 \log n)$ 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 Netherlands Organisation for Scientific Research (NWO) is supporting B. Speckmann under project no. 314.99.117. Castermans under project no. 639.023.208, K. Verbeek in the humanities.

patterns in such rich datasets that may be of great interest to researchers. Techniques will be required to facilitate the discovery of interesting painting that is moved between several museums. New visualization techniques will be required to address 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