Paper: Can Interactive Map-Based Visualizations Reveal Contexts of Scientific Datasets?

Eva Fischer  
The University of Western Ontario  
Faculty of Information and Media Studies

Olha Buchel  
The University of Western Ontario  
Faculty of Information and Media Studies

Abstract

Existing map-based visualizations of scientific datasets support a small number of tasks. They do not allow users to visually inspect properties and contexts in scientific datasets and focus only on showing locations in space and time. This paper describes a prototype that provides a better support for visual analyses of scientific contexts by means of additional representations and richer interactions with scientific data.

Résumé

Les techniques de visualisation actuelle par carte des ensembles de données scientifiques permettent un petit nombre de tâches. La principale raison est que la visualisation ne représente pas toutes les propriétés des ensembles de données. En visualisant uniquement des points à un temps et à un moment précis, une telle technique de visualisation a des capacités limitées aux fins d’analyse visuelle des contextes des ensembles de données scientifiques. La visualisation des autres propriétés peut améliorer l’analyse visuelle des contextes scientifiques. L’approche proposée est illustrée avec un prototype de visualisation.

Introduction

Map-based visualizations (MBVs) and illustrations are becoming popular representations of scientific datasets. They can be found in practically any data repository providing access to scientific datasets. The Creative Commons database PANGAEA\(^1\) and the Library of Congress’s *Biodiversity Heritage Library*\(^2\) are examples of repositories with MBVs. Their visualizations, however, can be considered more like illustrations (not visualizations) because they simply represent locations of samples described in publications. In order to qualify as visualizations in the information visualization sense, maps should be enhanced with widgets, controls, filters, and other tools allowing users to perform small actions on visualized data (e.g., filtering, selecting, repicturing, and so on) that may allow users to reveal interesting patterns in data.

The lack of interactions can be explained by the fact that visualizations largely overlook the potential role of ontological properties captured in datasets. The existing visualizations at the PANGAE A website and *Biodiversity Heritage Library* visualize only locations of samples and...
ignore all other properties of samples. Each dataset has many other ontological properties, which are used in many publications and their supplementing datasets. The expanse of properties provided by each sample has the potential for designing interactive MBVs that can expand visual analytical tasks of researchers.

In this paper we explain how researchers could benefit from MBVs if they were augmented with interactions and additional representations of ontological properties. We do this on a prototype dataset targeting dinoflagellate cysts. This and other similar datasets are the result of work done during Integrated Ocean Drilling Program and its predecessor missions. The program involves scientists from around the world to examine and interpret the geological significance of the ocean floor. Research from samples gathered during expeditions organized in the context of these programs have helped answer questions about Earth’s history and structure, the process of climate change, the use and potential of natural resources, etc.

Currently on the PANGAEA website there exist 1532 datasets that include dinoflagellates. While the researchers are able to collect these datasets, their ability to analyze datasets on MBVs without reading full text of publications lags behind mainly due to the lack of support for visual analytical tasks. For example, current visualizations prevent understanding of the following: How many samples do they describe? How samples in one paper are different from samples in other papers? This shortcoming can be overcome by visualizing ontological properties in datasets that supplement publications about dinoflagellates.

The paper is organized as follows: in the Background section we examine existing visualizations of scientific publications and explain our theory for the design of MBVs; in the Methodology section we present our dataset, describe our prototype visualization, and explain the advantages and significance of our approach.

Background

In information visualization much of the research focuses on visualizing scientific papers, but not datasets. For this reason, the majority of these visualizations show bibliographic properties and citations (see, e.g., Aikens, Lucchese, Webster, & Kerne, 2011; Yan & Ding, 2011a; Yan & Sugimoto, 2011b; Leydesdorff & Persson, 2010). Such visualizations facilitate understanding, reasoning, and decision making based on bibliometric analyses, but they overlook scientific contexts in publications.

In this paper we focus on designing visualizations for analytical reasoning about contexts in scientific papers. Visual analytics visualizations are being increasingly used in Earth sciences (White, 2008). Such visualizations combine automated analysis techniques with interactions for an effective understanding, reasoning and decision making on the basis of very large and complex datasets.

Our approach to conceptualizing visual analytical tasks and the strategy for supporting them builds on the research about visual tasks in geovisualization by Peuquet (2002), Andrienko and Andrienko (2007) and Andrienko et al. (2003). As detailed in that research, questions posed by space-time-attribute data analysis usually involve three components: where (space), when (time), and what (attribute/thematic objects) (Peuquet 2002). Each of these components should be
visualized. Andrienko et al. (2003) also introduced two “search levels” to the analytical tasks: (1) an elementary level in which a task deals with individual objects (such as a time, a place or a property); and (2) a general level in which a task considers a set of objects as general situations. Hence, researchers working with complex datasets need to first gain an overview of the entire data set to quickly understand the scope and structure of the dataset and discriminate between interesting and uninteresting content, then focus on subsets of data with more viable patterns in detail views. The support for these tasks largely depends on the use of additional representations (Buchel & Sedig 2011) and interactions (Buchel & Sedig, in review).

Methodology

To investigate the virtue of interactive MBVs, we are developing a prototype visualization that shows samples of dinoflagellates described in the datasets on the PANGAEA website. Specifically, we took 1 dataset with 204 samples. This dataset, from De Schepper et al. (2011), actually contains information regarding four distinct sites from the North Atlantic, located in close geographic proximity.

All samples in this dataset have identical properties. These properties describe chemical composition (specifically $\delta C^{13}$ and $\delta O^{18}$) as well as the dinoflagellate cyst content. Furthermore each sample has a depth measurement, information about correlation, and an age estimate. Although multi-proxy studies, looking at numerous biological proxies at the same time, such as those combined in De Schepper et al. (2011) are new to the field, dinoflagellate cyst datasets generally have the same ontological properties, e.g. depth, age, relative dinoflagellate cyst content, total dinoflagellate cyst abundance, pollen/dinoflagellate cyst ratio, etc. The total number of properties examined in the prototype is 68, focussing mainly on the relative dinoflagellate cyst content. The dataset was downloaded into an experimental MySQL database. The visualization prototype is being designed with Google Maps API, PHP and AJAX. In the remaining part of this section we explain the design of additional representations, interactions, and possible tasks that can be accomplished with the prototype.

Representations. The visualization uses three kinds of representations: a map, a pie chart, and a diagram. All sample locations are plotted on the map. Figure 1 shows geographic distribution of four different sites from the North Atlantic published in De Schepper et al. (2011). Although currently the representation has only four locations, it could be conceivable to plot all other
locations contained within the PANGAEA database.

Figure 1. Geographic representation of datasets from the North Atlantic published in De Schepper et al. (2011)

Pie charts (see Figure 2) will be used for representing dinoflagellate abundances at each depth level. The use of the pie charts allows for a quick visualisation of everything that has been measured at a specific depth within the specified area examined.

Figure 2. Pie chart of relative dinoflagellate abundances constructed from the dataset of De Schepper et al. (2011) at depth 14.1m
The third type of representations will be a stratigraphic diagram presented in Figure 3 that allows overviewing properties of cysts from each individual site. Such diagrams allow researchers identify the relationship between dinoflagellate cyst abundances and depth. Depth is displayed along the y-axis for each species, and each species is thus displayed side by side.

Figure 3 -- Stratigraphic diagram for the main dinoflagellate cysts found at DSDP Hole 603C (one of the sites included in the De Schepper et al. 2011 dataset) (Fischer, unpublished thesis)

**Interactions.** The prototype visualization has several interactions that allow users to perform visual analyses. These interactions are linking, filtering, and selecting. Linking connects dinoflagellates’ representations (pie charts) with the map. The pie charts are included in markers’ information windows. Filtering allows acting on samples’ representations by adding and removing samples from the map. Figure 4 shows 2 continuous and a number of discrete filters. The continuous filters allow controlling the number of samples in view by depth and time; the discrete filters allow selecting the results that fit certain criteria.
And finally the selecting interaction allows drawing a minimum bounding box around markers on the map as shown in Figure 4. Once the markers are selected, samples from them are represented as a stratigraphic diagram, shown in Figure 3.

**Tasks.** These additional representations and interactions can significantly expand a number of tasks that can be performed with interactions. They can facilitate comparing distributions of samples in space, depth, and time; exploring relationships between various species; and consequently it may lead to discovery of commonalities and anomalies.

**Significance**

The proposed approach can help researchers better understand scientific data about dinoflagellates. It can complement one of the palynologist’s main tools, *The Lentin and Williams Index of fossil dinoflagellates, 2004 edition* by Fensome and Williams (2004). While the publication lists all formal and semi-formal names of dinoflagellate cysts, the holotype and its stratigraphic range, the visualization can show how dinoflagellates interact in space and depth. Finally, if this methodology proves to help researchers make interesting findings, it may imply that additional representations and interactions may enhance visual analytics in contexts of many other datasets.

**References**


1. http://www.pangaea.de/
2. http://www.biodiversitylibrary.org/browse/map