WEB-BASED VISUAL EXPLORATION AND ERROR DETECTION IN LARGE DATA SETS: ANTARCTIC ICEBERG TRACKING DATA AS A CASE

Connie A. Blok
blok@itc.nl

Ulanbek Turdukulov
turdukulov@itc.nl

Barend Köbben

Juan Luis Calle Pomares

International Institute for Geo-Information Science and Earth Observation (ITC)
P.O. Box 6
7500 AA Enschede
The Netherlands

Abstract

Polar iceberg data are – amongst others – used by scientists who are interested in complex phenomena like global warming, climate change and changing habitats. Other main users of the data are organisations and individuals involved in navigation and engineering that need to take iceberg positions and behaviour into account for their activities. Iceberg positions are identified, tracked and, together with some other attributes, made available on the Web. Data tracking and storage are manually performed, and therefore error-prone. If these errors are not detected and handled, they will propagate and negatively influence the quality of the derived information.

The main objective of this paper is to demonstrate that an interactive environment to visually explore the characteristics of spatio-temporal data can *also* be used to detect errors in the data set. We report on the method used to design and implement a web-based environment to visually explore iceberg characteristics. Although the environment was not intended for error detection, we found that several errors in the data could be quickly discovered. Main limitations in the current propotype are described, together with recommendations for improvements.

Introduction

A broad research community is interested in the distribution and behaviour of icebergs. The research is stimulated by phenomena like global warming and climate change, environmental problems like changing habitats, and navigation and engineering

activities that are endangered by floating or grounded icebergs. The recently ended Fourth International Polar Year (2008) stimulated many research activities and resulted in numerous publications and attention of public media. Among the characteristics of icebergs that are studied are their spatio-temporal distribution, movement dynamics, and events like calvings: splittings from iceshelves and from each other.

The data are collected using different enhance resolution scatterometer and radiometer instruments, but the detection of icebergs, and particularly their tracking, is a difficult process (Ballantyne and Long, 2002). Main reasons are that wind and melting conditions (particularly in summer) cause big changes in the brightness and contrast of both icebergs and the surrounding sea ice and ocean. Various algorithmic processes have been tried, but so far none of these provides consistent results due to the variable and dynamic conditions under which the tracking takes places. Recent altimeter-based techniques have improved the detection in open water, but that is not enough yet, and they cannot track (Tournadre et al., 2008). Tracking, therefore, is still a manual process, based on human interpretation and therefore error-prone.

Tracked data (from 1976 to present) are made freely available by the US National Ice Center (NIC), responsible for naming and operational monitoring of icebergs around the world (*URL 1*), but the data files contain many errors. These errors are partly related to the tracking problems sketched above, but there are also other human-induced problems, like typing errors. These errors seriously limit the use of the data, and if errors are not detected and properly dealt with, they will propagate and hence influence knowledge extraction. This problem becomes even more severe if the iceberg data sets are used to search for (possibly causal) relationships with other phenomena, e.g. in global warming or climate change studies.

Focus of the project

Currently, we are conducting research in several areas that are related to – or make use of – the iceberg data set (see also (Turdukulov and Blok, 2008). Here we will report on first results obtained in one of our studies (Calle Pomares, 2009). Given the broad interests in iceberg data from distributed users in various scientific fields, who may want to combine iceberg data with other data, a main objective of this study was to provide an interactive, web-based environment to visually explore the characteristics of the iceberg data.

At the start of the project, we found some existing web-based visualizations of Antarctica and surroundings, but they have limitations that vary from no focus on icebergs, or no representation of iceberg trajectories, to no time stamps or only covering limited time spans with recent icebergs, and/or limited interaction, particularly with the temporal component of the data. Some examples are given here, but no attempt was made to give an exhaustive overview of web-based visualizations, and developments in this area are fast (*URL 2-5*).

Another main objective of the study was to investigate the feasibility of interoperable SIMILE Tools for the development of such an environment. Although there are many other visualization libraries (see e.g. *URL 6*), the main reason for the focus on SIMILE tools for the prototype was that the temporal component is important for tracked data, and Timeline is one of the tools that offers display of, and interaction with, the temporal component of the data together with a spatial view. There was, however, no need to limit the project to the use of SIMILE tools only. SIMILE (acronym for Semantic Interoperability of Metadata and Information in unLike Environments) is a research

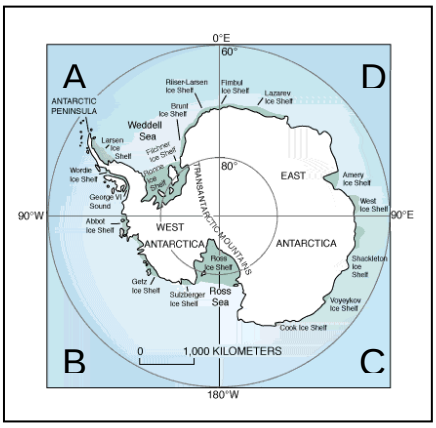

Figure 1. Naming of icebergs is based on quadrants of origin (adapted from: www.solcomhouse.com/images/antarcmap.gif)

project conducted by MIT; its aim is to develop tools to increase the interoperability among distributed (meta)data, services, etc. (*URL 7).*

The intention of the above mentioned study was *not* to include means for error detection It turned out, however, that the environment that has been created supports quick error detection. The main objective of this paper, therefore, is to demonstrate that an environment to visually explore the characteristics of spatio-temporal data can *also* be used detect errors in the data set.

Processing and analysis

In the project mentioned above (Calle Pomares, 2009), a subset of the NIC data archives for Antarctica (the area located south of 60º latitude) was used. The subset included icebergs with positions recorded between 2000-2007. Most of the time, recordings were made every 1-5 days. One of the attributes in the files is the name of the iceberg. Naming occurs according to the quadrant of origin (figure 1), followed by a sequential number. When an already identified iceberg breaks apart, its subdivisions are named by adding an alpha suffix to the name of the 'parent' iceberg, so A-23A and A-23B are

subdivisions of A-23. Other attributes in the NIC data set refer to positions (in space and time), the size of the icebergs, and the source (satellite).

Inspection of the data files revealed a few errors. In some records, either date, location, or both were missing. In the first two cases, missing attribute values could be estimated by simple interpolation based on neighbouring space-time positions. Records with missing date *and* location had to be removed, just like duplicates. For partial duplicates – two records with identical names and dates, but different positions – it was not so easy to determine which one should be removed. The decision was supported by plotting all icebergs onto Google Earth. Based on visual inspection and the assumption of proximity with respect to previous/next recorded positions, likely erroneous duplicates could be removed. Further visual inspection, however, also revealed strange, out of context positions along the trajectory of some icebergs. Some of these clearly locational errors could be corrected because they could easily be traced back to typos in the database (exchange of plus/minus signs for North/South or East/West; exchange of latitude/longitude). If no cause could be found, errors were replaced by interpolated estimations of positions.

Web-based visual exploration of the distribution, movement characteristics and events related to icebergs and possibly other multi-dimensional data needs an environment that offers at least interactive, dynamically linked spatial, temporal and multi-dimensional attribute views on the data. Those views should, according to the Information-Seeking Mantra propagated by (Shneiderman, 1996) provide *overview* of whole data set, enable *zooming* and *filtering* to focus on portions of interest and exclude things that are not relevant in the current context, and provide *details on demand*. These broadly defined requirements were used to guide the design and implementation of the prototype.

Design and implementation

Design started with generating some additional attributes, derived from the ones provided by NIC. Among them are: lifetime (existence from origin to calving or dissolution/loss), total distance travelled by each iceberg, speed (between recorded positions and average speed over the whole trajectory), orientation of movement (azimuth), and area status (reduction or no change). Some of these attributes were used to enable better filtering in the prototype, others were displayed in information pop-ups.

As mentioned above, some SIMILE tools, and particularly the Timeline, were considered to be a potentially relevant for the prototype. SIMILE tools (*URL 7*) form a library of about 20 tools: client-side JavaScript-based Web API's, supported by AJAX technology, thus enabling integration of heterogeneous resources and formats. SIMILE tools are supposed to be slim, fast, and as cross-browser platform as possible.

Among the tools is Exhibit 2.0**,** a framework that enables authors to create interactive 'exhibits' of data collections in an interface, and publish it on a web server without a

need to rely on database and server-side technology. Exhibit has been developed by (Huynh et al., 2007). Many web-based tools can be embedded, also from other libraries. Exhibit supports data visualizations, browsing and progressive filtering, and has been used in this project to build the user interface.

Exhibit includes a Map View extension that interacts with the Google Maps API (*URL 8)*, so Google Maps can be embedded to provide a spatial view (using the map, image or hybrid layer). Authors can manipulate the view and add content. But the representation is static, with Antarctica spread all over the bottom, and in the then available customized version of the Google Maps API 2, the path of an iceberg could only be visualized by a series of points.

Figure 2. Lifetimes of icebergs are mapped on an interactive Timeline. The lower part provides overview over a number of years, while the upper part shows details for January-May 2003.

With the *desktop* version of Google Earth, on the other hand, users can easily pan and rotate to get a good view on the Antarctic region, there is no projection problem. Furthermore, time stamps can be added and dynamics be viewed using the Timeline; kml files and other content can easily be loaded, and display of trajectories is no problem. Google Earth can also be *embedded* in web pages through a plug-in and JavaScript API (*URL 9*), but the version used in the prototype (1.002) could neither directly be integrated in Exhibit's Map View, nor display animated trajectories (see the section 'Main results' below). Google Maps and Google Earth, being partly complementary, were both selected in the project described here.

Timeline (originally developed within the SIMILE project) is a tool to map time-stamped data onto an interactive widget. It helps users to browse through temporal data, make selections and query the data. It is a flexible and customizable tool that enables the generation of different timelines, including views that have different bands that can be configured to provide overview and detail. The bands can be synchronized so that panning in one band also scrolls the other. It is also possible to embed links or images. Timeline has been selected for the representation of the temporal component in the

project (figure 2). Another view, selected to explore relationships between pairs of attributes, was SIMILE Scatterplot.

To provide an overview of the multi-dimensional attribute space, the option to include tiles in the interface was used. Tiles enable faceted browsing: users can navigate in the multi-dimensional information space and progressively narrow their choices in each dimension by filtering. In the project, filtering is enabled on iceberg ID, year and month of appearance, lifetime, average speed, area, direction of movement and on latitude/ longitude intervals. Values of some of these attributes have been grouped into a number

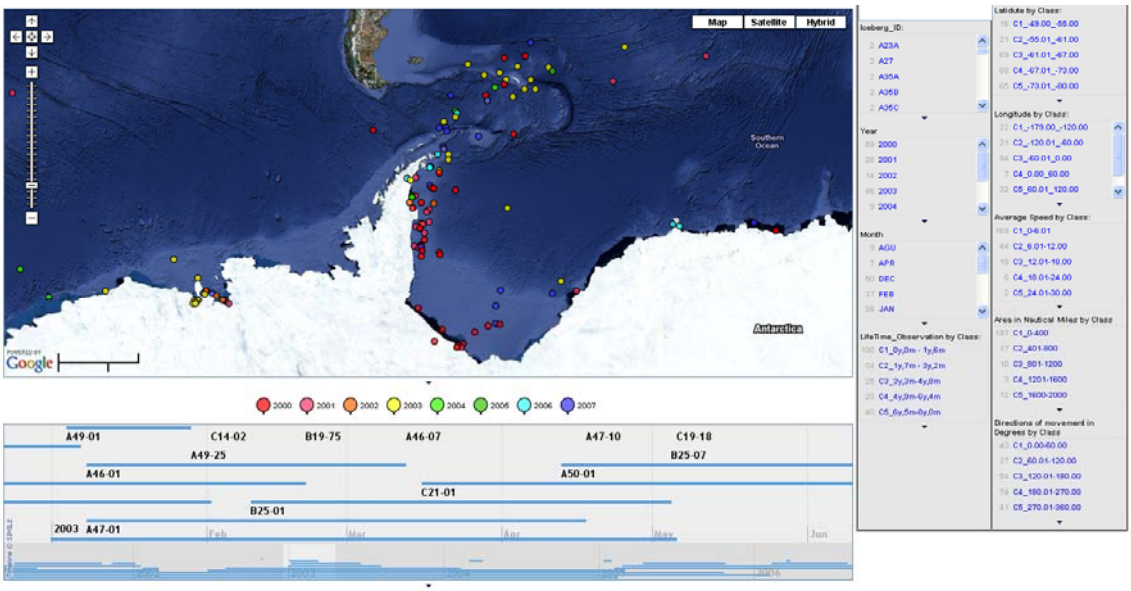

Figure 3. Exhibit interface with Map View, Timeline and facets; provides overview and enables zooming, filtering and getting details on demand.

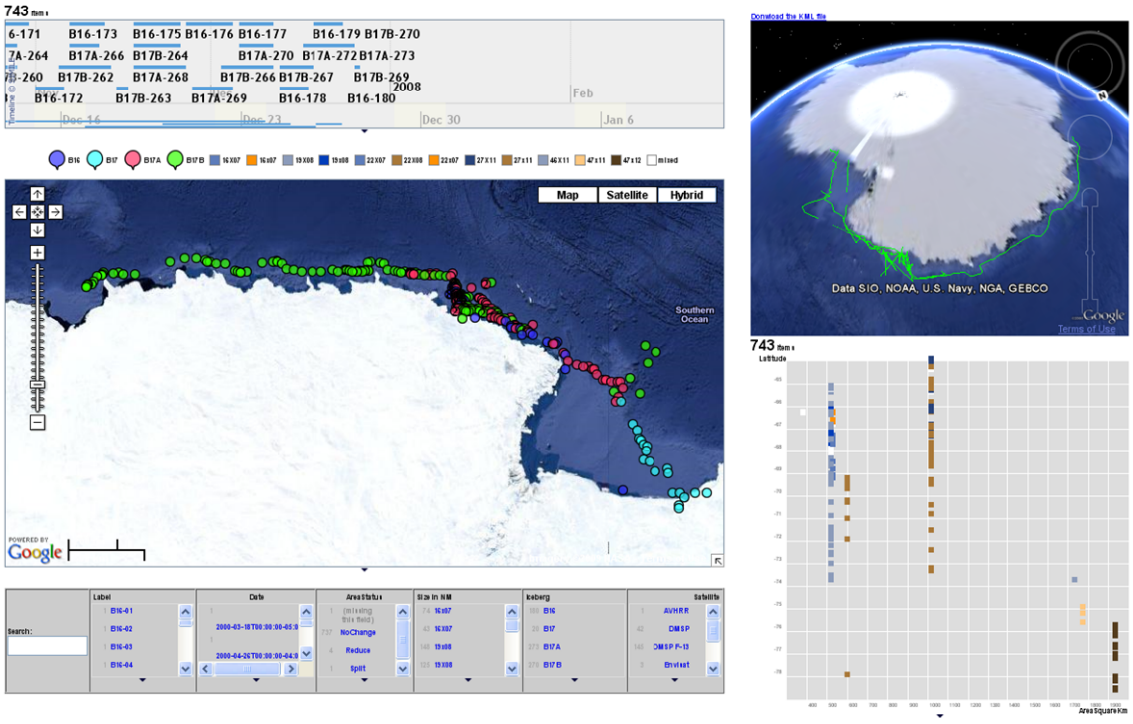

Figure 4. Separate window to display trajectories in Google Maps and Google Earth, with a search pane and Timeline. Furthermore, a Scatterplot has been added.

of classes, and filtering can only be performed on those classes; in other cases filtering is possible on individual values.

All the components mentioned above have been integrated in the prototype. Figure 3 shows the main window. When elements are selected in the Map View or the Timeline (figure 2), information pop-ups list all the elements' attributes, and if a user selects an iceberg, its trajectory is displayed in a new window containing Google Maps, Google Earth and also the Scatterplot (figure 4).

Main results

A research propotype has been developed that consists of several linked data views (like a Map View, Timeline, Scatter plot) and functionalities in a user interface. The prototype is meant for visual exploration, and should provide users with overview and options to zoom, filter, and view details on demand.

We have noticed the following limitations of the current prototype:

- Its input is limited to 10,000 records. The prototype has dealt with 9,516 iceberg records corresponding to 118 icebergs for the period 2000-2007, but if the limit is exceeded, the browser's performance slows down and synchronization of elements in Exhibit's Timeline and Map View becomes unresponsive. The limitation can be

avoided by aggregating input data according to different zoom levels. In the current implementation, this has been realized, but not in an optimal way: access to different levels is attained via a link option that sends the user to another web page.

- Some main limitations of Google Maps (static representation, projection that cannot be changed, e.g. into polar stereographic centered on Antarctica) can be avoided by Google Earth, but the Google Earth API - designed to provide Google Earth functionality on web pages - cannot be directly integrated in Exhibit's Map View due to some incompatibilities. This limitation has been partly solved by including the Google Earth API in a new window that opens if a user selects an iceberg ID in the main window. The trajectory of an iceberg can then be seen in 2D and 3D, but the Google Earth API used in the prototype is still a beta version that has limitations compared to its desktop counterpart, such as absence of a Timeline, hence animated trajectories cannot (yet) be displayed. The earlier mentioned problem that movement paths could only be represented by series of points in Google Maps, and not by lines has been solved in the recently released newer version of the Google Maps API.
- An important limitation refers to dynamic updating and storage of data in the database on the server-side. Current input to the prototype is realized via flat (json) files. Thus classification of some iceberg attributes in the facets is hard written into the json files and there is no mechanism that can update these classifications on the fly. It limits querying and filtering processes, and makes updating cumbersome. This limitation can be solved by using server-side support and migrating iceberg data to a database.
- The prototype is not compatible with some browsers: Safari and Opera browsers cannot run the Google Earth API; Internet Explorer cannot correctly display the main Exhibit page. There are no compatibility problems in the Firefox browser.

Although we tried to clean the data files before visualization, we found not only that visualizing the data in a map reveals additional errors (see 'Processing and analysis'), but we also discovered that by filtering and drilling down using the prototype, additional problems and errors, like unrealistic, extremely high speed of movement or weird trajectories, could quickly be found. These are typically problems that do not show easily by inspection of the database. If these kind of problems can quickly be detected, particularly in large spatio-temporal data sets, user-guided removal of errors will improve the quality of the data set, and the potential to derive better conclusions.

Conclusions and recommendations

Main objectives of the initial project (Calle Pomares, 2009) were to provide an interactive, web-based environment to visually explore the characteristics of the iceberg data, and to investigate the feasibility of interoperable SIMILE Tools for the development of such an environment. Visual exploration of the iceberg data is currently not yet fully supported in the prototype. Mapping data onto the SIMILE Timeline reveals temporal patterns and making use of facets provides interesting opportunities to drill down into the data. But spatial and spatio-temporal patterns are difficult to perceive

from a static representation of the first and last observation of an iceberg only, or even from static representations of the trajectories. No behaviour and events (like calvings, size changes, accelerations) are dynamically represented. The feasibility of SIMILE Tools would improve if the Google Earth API with an option to add animation could be included in Exhibit's Map View. Options to add other dynamically linked views would even further enhance pattern comparison and discovery of relationships (e.g. a scatterplot matrix, a Parallel Coordinate Plot).

Another objective of this paper was also to demonstrate that an interactive environment to visually explore the characteristics of spatio-temporal data can *also* be used to detect errors in the data set. We have provided examples of errors that were not revealed by inspection of the data records, but that could easily be detected after visualization and drilling down into the data (e.g. strange locations, weird trajectories and extreme high speed of movement). Such results can be used to better preprocess and improve the quality of the data. Currently, we are in a transition phase between manual and automated procedures to collect and preprocess many types of geodata; ultimately the manual processes will vanish, but for the time being, use of a web-based visual exploration environment that supports finding of problems seems a useful approach. Offering such a platform as a *Web Service* is an advantage for the research community that is interested in iceberg patterns and dynamics. Users would be able to display newly added data to the tracking data base, and - if needed - correct or disregard records before the data are further applied by the users for their own purposes.

The project described here was executed at the end of 2008 till early 2009, but web developments are fast. We have seen some improvements already, and there are no doubt many effective tools available. Further research will focus on modeling, database support, server scripting, supplementary tools, additional functionalities and better visualization of data. So far, we used icebergs as a case, but results should be broader applicable, particularly to applications using trajectory data.

Lastly, the project did not involve users in the design of the prototype. User-centered design (van Elzakker and Wealands, 2007) would be recommended to create an effective, consistent and user-friendly Web Service.

References

BALLANTYNE, J. & LONG, D. G. (2002) A multidecadal study of the number of Antarctic icebergs using scatterometer data. *Geoscience and Remote Sensing Symposium, 2002. IGARSS apos;02. 2002* IEEE International.

CALLE POMARES, J. L. (2009) Visualizing iceberg movement data using 'SIMILE' AJAX framework. Enschede, ITC.

HUYNH, D., F. , KARGER, D., R. & MILLER, R., C. (2007) Exhibit: lightweight structured data publishing. *Proceedings of the 16th international conference on World Wide Web*. Banff, Alberta, Canada, ACM.

SHNEIDERMAN, B. (1996) The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. *Proceedings of the IEEE Symposium on Visual Languages.* Washington, IEEE Computer Society Press.

TOURNADRE, J., WHITMER, K. & GIRARD-ARDHUIN, F. (2008) Iceberg detection in open water by altimeter waveform analysis. *J. Geophys. Res.*, 113.

TURDUKULOV, U. D. & BLOK, C. A. (2008) Visual analytics to explore iceberg behaviour: extended abstract and powerpoint. *Presented at Workshop 'GeoVisualization of Dynamics, Movement and Changeâ€™ of the ICA Commission on GeoVisualization at the AGILE 2008 Conference, May 5, 2008. Girona, Spain. 4 p. + 16 slides.*

VAN ELZAKKER, C. P. J. M. & WEALANDS, K. (2007) Use and users of multimedia cartography. *In: Multimedia cartography. / ed. by W. Cartwright, M. Peterson and G. Gartner. Second edition. Berlin : Springer, 2007. ISBN: 3-540-36650-4. pp. 487-504.*

URL's

1: US National Ice Center: current iceberg positions and archives <http://www.natice.noaa.gov/products/iceberg/>

2: National Snow and Ice Data Center, Atlas of the Cryosphere <http://nsidc.org/cgi-bin/atlas_south?layer=sea_ice_extent_01&layer=snow_extent_01&layer>

3: British Antarctic Survey, Antarctic Digital Database <http://www.add.scar.org:8080/add/WMSmap.jsp>

4: US Antarctic Resource Center, Atlas of Antarctic Research <http://gisdata.usgs.gov/website/antarctic_research_atlas/viewer.htm>

5: US National Ice Center, Current Antarctic Iceberg Positions <http://ice-kml.natice.noaa.gov/>

6: A Beautiful Web, Todd Holloway's Visualization libraries <http://abeautifulwww.com/2008/09/08/20-useful-visualization-libraries/>

7: SIMILE project of Massachusetts Institute of Technology: tools Wiki <http://simile.mit.edu/wiki/Main_Page>

8: Google Maps API <http://code.google.com/intl/nl/apis/maps/>

9: Google Earth API <http://code.google.com/intl/nl/apis/earth/>