SFMN GeoSearch: An interactive approach to the visualization and exchange of point-based ecological data

Rodolphe Gonzales, Jeffrey A. Cardille, Lael Parrott, Caroline Gaudreau, Gaël Deest

Département de Géographie, Université de Montréal, Montréal, QC, Canada
École de Technologie Supérieure, Montréal, QC, Canada
Département de Mathématiques, Université d'Angers, France

Please cite this article as: Gonzales, R., et al., SFMN GeoSearch: An interactive approach to the visualization and exchange of point-based ecological data, Ecological Informatics (2009). In press. doi:10.1016/j.ecoinf.2009.07.007

Keywords: Data sharing, Geographic web portal, Virtual globes, Geobrowser, Spatiotemporal visualization.

Abstract

Recent advances in computer networks and information technologies have created exciting new possibilities for sharing and analyzing scientific research data. Although individual datasets can be studied efficiently, many scientists are still largely limited to considering data collected by themselves, their students, or closely affiliated research groups. Increasingly widespread high-speed network connections and the existence of large, coordinated research programs suggest the potential for scientists to access and learn from data from outside their immediate research circle. We are developing a web-based application that facilitates the sharing of scientific data within a research network using the now-common “virtual globe” in combination with advanced visualization methods designed for geographically distributed scientific data. Two major components of the system enable the rapid assessment of geographically distributed scientific data: a database built from information submitted by network members, and a module featuring novel and sophisticated geographic data visualization techniques. By enabling scientists to share results with each other and view their shared data through a common virtual-globe interface, the system provides a new platform for important meta-analyses and the analysis of broad-scale patterns. Here we present the design and capabilities of the SFMN GeoSearch platform for the Sustainable Forest Management Network, a pan-Canadian network of forest researchers who have accumulated data for more than a decade. Through the development and dissemination of this new tool, we hope to help scientists, students, and the general public to understand the depth and breadth of scientific data across potentially large areas.

1. Introduction

Independently accumulated datasets are, for the most part, bound to personal computers in laboratories, while an aggregation of these individual efforts could be used to address the complex ecological questions with which humanity is confronted. In ecology particularly, scientists are often working on problems that need to be tackled from a multidisciplinary perspective with data covering wide spatial and/or temporal spaces (Michener et al., 1997; Bowker, 2000; Andelman et al., 2004; Carpenter, 2008). Although collaborations between specialized researchers from different disciplines are frequent, once a study is completed by a research team, results are typically
communicated through the usual peer-reviewed articles, publications and conferences, and the data that led to the published results are seldom released outside of the laboratories from where they emerged. The existence of these disconnected individual datasets creates a situation where bits of information about a given subject are geographically and/or socially scattered within the research community, and therefore not exploited to their full potential.

Over the last 15 years, the issue of aggregating and sharing ecological datasets has been a very active domain of research. Ground breaking works have been, and still are, pushing towards freeing the data's potential by providing scientists with better ways to tackle challenges related to gathering, documenting, organizing, and redistributing spatially and temporally scattered heterogeneous datasets for second-hand uses (Michener, 2006). In this context, tools to set up data sharing frameworks like Metacat (Jones et al., 2001), or ecology-oriented online systems based on aggregated databases, like OBIS-SEAMAP (Halpin et al., 2006), Ecotrends (Servilla et al., 2008) and GBIF-MAPA (Flemons et al., 2007), are contributing to a new paradigm of broader data sharing (Guralnick et al., 2007).

Meanwhile, in the last 5 years, the emergence of new mapping methods, aimed at cartographers and non-cartographers alike, allow for novel ways to display large amounts of data on a geographical medium. These methods take advantage of the spread (at least in the richest parts of the world) of increasingly fast Internet connections, as well as of the progress made in consumer-grade computer graphics. Commonly called “virtual globes”, or “geobrowsers”, they provide interactive three-dimensional models of the Earth on which all sorts of content can be downloaded from remote servers and shown in quasi real-time. Virtual-globe applications like Google Earth (earth.google.com) and NASA World Wind (worldwind.arc.nasa.gov) are popularizing this new approach and greatly expanding humanity's options for interacting with geographic information. The scale and speed of their adoption for a wide range of uses is striking: more than 100 million users downloaded Google Earth in its first year, and many millions of user-created maps have been made using Google Maps technology. For an extremely broad audience of academics, stakeholders and general public, virtual globes provide what might be considered a simplified, introductory Geographic Information System. Though virtual globes do not currently provide analysis capabilities as sophisticated as those of their professional GIS counterparts, they are broadly considered to be easy to use, highly interactive, and capable of inducing a good understanding of ecological data through their geographic context (Tooth, 2006; Foresman, 2008; Goodchild, 2008).

With the growing interest in data sharing, recent years have seen a variety of initiatives taking advantage of these advances to visualize georeferenced shared scientific datasets. Within this new cartography, two kinds can be distinguished: two-dimensional, projected maps, and three-dimensional models of the earth. Projects like OBIS-SEAMAP (Halpin et al., 2006) are good examples of the frontier of mapping publicly available ecological databases in two dimensions: at a given website, users make queries and the resulting datasets are displayed on a simulated projected map of the Earth. More recently, researchers have begun to use three-dimensional virtual globes as a geographic visualization medium. Many projects employing virtual globes are built around Google’s Keyhole Markup Language (KML) format, an increasingly popular XML-based file format originally designed around Google Earth. KML files can contain the geographical positions of points, lines, polygons or images in order to build and exchange preset data visualizations on virtual globes. Users generally download a KML file pointing to a small number of data items (either embedded in the KML or stored remotely) for display in Google Earth or on other virtual globes. In turn, users can create a KML out of their own data, exchange it with other users, and/or submit it to a remote server. The KML language is an Open Geospatial Consortium standard (http://www.opengeospatial.org/standards/kml/) and its simplicity, flexibility, and openness are leading to its broad adoption across platforms and applications. Systems using an external virtual globe to display queries made in another application facilitate knowledge sharing by presenting different datasets in their spatial context. But unlike data sharing tools using two-dimensional cartography, they still require users to put a set of different tools to work in order to go through the process of searching and displaying
data. The lack of integration in the process of exploring data and visualizing them could be, for many scientists, an obstacle to the use of geographically explicit, shared data exploration systems.

We present here the latest progress made on SFMN GeoSearch, a data sharing and visualization application built for the Sustainable Forest Management Network (SFMN). The SFMN is a Canada-wide research group connecting individuals from heterogeneous sectors of activity and promoting interdisciplinary collaborations in research related to the sustainable management of forests in Canada. It has been a member of Canada’s Networks of Centres of Excellence since 1995, and has gathered over time more than 450 people active in a broad range of fields.

The focus of our research is to provide ecology-oriented research networks with a generic, user-friendly, integrated tool that takes advantage of the latest in three-dimensional mapping technologies. Instead of harnessing separate programs to search and to display information, we built an integrated system where the tools needed for database exploration, data visualization, and communication between datasets authors and potential second-hand users are tightly inter-connected, easy to use and accessible from the same online application.

In this article, we present the major aspects of the SFMN GeoSearch framework, focusing on its modules related to the spatiotemporal visualization and the exploration of shared datasets.

2. **Structure of the application**

The SFMN GeoSearch system provides a set of heterogeneous modules and techniques, each playing a distinct role in data storage and visualization (Fig. 1). These tools, integrated in SFMN GeoSearch as a single interface allow members to: (1) explore in intuitive ways the sum of all data already submitted; (2) visualize chosen data on a multidimensional geographic medium; and (3) discuss, annotate and comment on the submitted datasets. The architecture of the system can be divided in two main parts: a user interface on one side, and a user-populated database on the other. The interface was made from various open source Java libraries and encapsulated in a web application, making it compatible with most popular web browsers. The database was built on PostgreSQL (www.postgresql.org), an open-source database management system; it parses, organizes and stores georeferenced variables from the datasets in order to support our spatiotemporal visualization system.

3. **Exploration of the database**

SFMN GeoSearch’s database is accessed and explored through queries built by users. While traditional GIS mainly use SQL (Structured Query Language) to retrieve, analyze data and build geographic visualizations from a database, we chose to keep the queries simple from a user’s point of view. We designed two means to retrieve information from the database while keeping users connected to the context of their data exploration.

3.1. **Building queries through a hierarchical tree exploration**

The entire database can be browsed through a textual exploration in a novel and dynamic hierarchical tree view. The tree view, or tree browser, is built on the prefuse java library (Heer, 2004), which introduces powerful and intuitive ways to visualize information and to interact with it. It represents the content of the database through a hierarchy of categories. Categories closer to the “root” of the tree are more general, while the ones closer to the “leaf”, the last items of the tree, are more precise. The exploration can be done based on variables, projects, sites or authors of studies. As a category, or node, is selected, more branches expand to display what belongs under it. For instance, a user looking for all research done in the province of Québec would first select the node “Sites”, then “Countries”, then
"Provinces", and finally "Québec". All research located in Québec would then be made available. The process (four mouse clicks) is more intuitive than a SQL query. It also retains a better sense of context as a user jumps from more general categories to more particular ones while keeping a clear idea of his/her position in the data tree (see Fig. 1-c).

3.2. Filtering data according to their date

Considering the value of the temporal dimension in interpreting change and discovering patterns in ecological data, it is important that the exploration of the database include a way to organize data along a time scale. The database is populated with variables sampled at different places and at different times. Some sites have had the same variable
sampled many times on the same spot. Similarly, a variable taken from a variety of datasets sampled across a wide geographic range is unlikely to have been sampled on the same date. This makes exploring and displaying time sensitive variables problematic: without taking into account their date of sampling, visualizing different datasets simultaneously could potentially be misleading. Therefore, users do need to be able to explore the database and select variables to be displayed according to their temporal dimension. We address this issue by displaying at all times a dynamic time scale (Fig. 1-g) linked to the geographical visualization described in the next section. Every time a query is done and a visualization is processed, the time scale adjusts itself to show the range of dates related to the displayed data, from the oldest sample to the most recent one. Users then (1) have a precise idea of the range of time stamps related to the data they are dealing with, and (2) are able to narrow down the search to certain dates in order to view a variable's values in a given time range.

3.3. Finding and selecting data via a graph of their authors' social interconnections

Graphing social interconnections in a complex research network can be challenging. In a research network such as the SFMN, members are related to each other through a variety of characteristics: some may have worked together in the past, some may share the same research interests, and some may belong to the same sector or institution. Graph-based methods of data exploration help anchor people-oriented data searches in their social contexts, and provide a better general understanding of the social interactions within the community. Social sciences, in their efforts to understand human interactions, have developed methods to analyze and map social networks (Scott, 1991). These analysis methods, based on mathematical graph theories, have led to powerful network visualizations (also called sociograms) where nodes, representing each member of the network, are connected to other nodes when a particularity is shared. The result is a drawing of interconnected nodes that effectively maps a network of relationships and offers a clear presentation of the global structure of the social interactions at work between members.

New methods of interactive network visualization such as prefuse have recently emerged and can be used towards designing better mediums for understanding social networks, as well as to explore, discover and select people-related resources in a database. Working from the concept of the traditional sociograms, prefuse innovates by taking advantage of recent advances in computer graphics and computational speed to present dynamic and, most importantly, interactive views of social networks. Once a graph has been shaped by selecting the links that should represent the social connections in the network, users can interact with the drawing by zooming in and out, or by panning in all directions. This helps users focus on specific people and follow connections from one scientist to another (Figs. 2 and 3). In a research network such as the SFMN, where nearly each node is directly related to potentially valuable resources, this intuitive way to explore graphs proves itself to be: (1) highly suitable to locate and access people-related data of potential interest in order to share them among members (Schwartz and Wood, 1993). While alphabetically sorted lists of members falling in similar categories could technically provide access to data according to people's social connectivity, they lack the constant sense of context that makes exploring a social network, and discovering its underlying resources, as easy as reading a map; (2) useful for members of a dense, geographically scattered network to comprehend their personal connectivity to the rest of the community. Beyond one or two degrees of separation, it is hard to mentally map one's own social connectivity, let alone understand other members' surroundings. Having an interactive and organized map of the network gives users an overall view of the general structure of their community and, by doing so, promotes awareness of their social environment. (3) Illustrate the role that a large centre of excellence plays in connecting researchers within a field, by allowing comparisons of the network connectivity before and after the centre's existence.
Fig. 2. An example of a social network represented as a graph with prefuse. Members are connected with each other if they have worked on the same projects (green edges), or if they belong to the same institution (blue edges). Prefuse takes advantage of a force-driven layout system to find an equilibrium point where sub-groups, hubs, isolates etc. are made visible. This visualization helps finding data through their authors' social interconnections.
Fig. 3. Implementation in our system of the prefuse network visualization library. It allows intuitive explorations of the database via an interactive graph of people's interconnections.

4. **Spatiotemporal visualization of field data**

The new digital cartography applications available today feature immersive methods to visualize and interact with georeferenced layers of information. There are two groups within these new geographic visualization clients. Some of these applications, like Google Maps or Yahoo! Maps, use projected maps or satellite imagery on a flat, two-dimensional medium. Others, like World Wind, Google Earth or Microsoft Virtual Earth, feature a model of the earth, around which stitched-together satellite images of different resolutions are wrapped. These systems effectively free users from preset geographic visualizations by providing intuitive ways to interact with the cartographic medium. The user controls the position of his/her point of view with the computer mouse, as if manipulating a virtual camera from a
remote location in outer space. Similarly, it is easy to move further back to visualize large features of the globe, zoom in to get closer to a specific region in order to observe details through satellite images of finer resolution. Users can also roll, or tilt the sphere, explicitly showing the topography of the area, and allowing the exploration of an accurate model of the earth’s geography in both its horizontal and vertical dimensions. Indeed, these applications do provide, through the continuity of the maps and the liberty of movement, a more intuitive and higher level of spatial understanding. For these reasons, we believed such scale-free geographic exploration medium would be an important addition to our system, and we worked towards a tight integration of NASA World Wind’s (WW) virtual-globe technologies within our ecological database exploration tool. At the beginning of the project in 2007, several systems were offering a fully interactive three-dimensional visualization of the globe. However, only WW had made available their source code as a Java library, allowing us to compile the program as a web browser-compatible Java applet. Furthermore, unlike similar products offered by Google, Yahoo!, and Microsoft, WW’s code is open source, and does not require any data to travel through third party servers. We believe this is an important point as our system could potentially be dealing with sensitive data.

While WW easily displays data spatially, visualization of queries in the temporal dimension has to be addressed in order to provide important context information to users. In the context of a meta-database, where different people have sampled data at different times and at different geographical positions, a medium explicitly displaying their spatiotemporal contexts provides strong visualizations in term of the understanding it can bring to the user. Along with spatial information indicated by two geographic coordinates, sampled variables stored in the database carry a time stamp. This extra dimension relative to each and every sample allows us to show their relative temporal values along with their geographic position. Thus, while plotting a point on the surface of the map represents spatial information, time is represented through the vertical, third spatial dimension of the virtual globe. These three values (latitude, longitude, and time) use the three geometric dimensions available on the 3D model of WW. Displaying the values of the variables then requires other means of visualization.

4.1. Visualizing variables

Aiming for a simple and intuitive method to display quantitative variable values while showing their positions in space and time, we designed a multidimensional visualization system based on a set of coded icons. For each space–time location, users can select two variables to be displayed on the globe. In addition to showing these variable values, we designed icons that give information -when applicable- about the dispersion of the means of the two variables displayed. The system is thus capable of showing up to seven pieces of information for each data point: the geographic position in longitude and latitude, the time of sampling, two variables of choice and their standard deviation. While multivariable geographic visualization systems do exist (Kreuseler, 2000; Andrienko et al., 2003), we decided, for reasons of clarity, to restrict the display to two variables.

Although the virtual globe is capable of showing points, lines, and polygons, we limited our approach to point-based data. There are several reasons for this restriction. First, nearly all attributes of a single point can be encapsulated for input in a single comma-delimited line, with limited risk of confusion. Second, the temporal attributes of lines and polygons can be substantially more complex than those of points: for example, if trying to represent a polygon for land annexed to a given park, the system might need to track the park’s original boundaries, the boundaries of the extension, the fact that the border between the extension and the old park became obsolete at a given point in time, and possible inholdings that could further complicate assigning a single time value to a particular polygon. Third, existing GIS systems have extensive analysis tools designed to operate on lines and polygons, and we reasoned that seeing them on a virtual globe was not as fruitful as the main focus, individual data points.
We employed several of the most common ways to convey information using an icon: color, relative size, and shape coding. We designed a set of icons that can display one or two variables in two different situations: the “top” view showing variables in two dimensions, and, as described below, the “tilted” view, taking advantage of the virtual globe’s third dimension to display time values.

4.1.1. Displaying one variable

When only one variable is to be displayed, the application shows a set of icons positioned on the sites where the variables have been sampled. The variables’ values are indicated by their relative sizes. As a rule, the lowest value will always be assigned to the same icon size (about 10 pixels in diameter). Similarly, the highest value will always be allocated to the same icon size (30 pixels in diameter). All other values will be distributed along a linear scale between these two extremes. No matter what changes of view the user chooses to do (zooming, panning, etc.), these sizes stay the same and match a scale displayed on the left side of the view window, showing the two extreme sizes as well as their corresponding values (Fig. 4).

4.1.2. Displaying two variables

When two variables are selected, the color of the icon comes into play. Values are separated in ten classes along a linear scale, and matched to a scale of colors (in the following order, from lowest to highest: purple, dark blue, light blue, cyan, turquoise, green, light yellow, dark yellow, orange, and red). The lowest values are then automatically set to purple, while the highest ones are assigned to red (Fig. 5). If the value tied to the color is missing on some sites, a grey icon shows the size. If, on the other hand, the missing value is the one relative to the size, a set of fuzzy icons (lacking a clear border), of constant size (10 pixels) is used by the system (Fig. 6).

4.1.3. Mean values

As stated before, one of the advantages in keeping such a large database is the ability to display changes in variable values through time, for situations in which the same variables have been sampled several times at different dates on the same site. All these data would then share, if not the same time stamp, the same spatial coordinates. Instead of having icons layered one upon another, with the latest sample masking the other icons, we chose to automatically
calculate the mean and the standard deviation of these stacked values. The only icons displayed are then the average of all the values belonging to the same point in space. However, this implies that icons should also bear a visual cue giving users an idea of the dispersion of data around the calculated mean.

4.1.4. Visualizing standard deviation

We need two standard deviation cues: one for the mean related to the variable represented by the size of the icons, and one for the mean related to variable represented by the color of the icons. Once again, trying to find solutions that would not require users to spend a long time observing the screen, we chose clarity over detailed value representation. For the value relative to the size of the icon, a 50% transparent stroke is applied to the icon; the thickness of the stroke corresponds to the standard deviation. For the color-related values, a half-disc of which the transparency is relative to the value of the standard deviation, is superimposed to the icon. Each of these visual representations falls into four classes (Fig. 7). For each icon, the choice of the class is relative to all the other standard deviations visualized at the same time on the virtual globe. In other words, at a given moment the lowest standard deviation values displayed are placed in the lowest cue class, the highest values are placed highest. The other standard deviation values are shown between them at regular spatial intervals.

Fig. 5. Visualizing two variables per site: one through the color, the other one through the size of the icons.

Fig. 6. Fuzzy icons show variables linked to colors only, with no size information.

Fig. 7. Icon A shows data that haven't been averaged or aggregated. Icons B to E show increasing values of standard deviation related to variable linked to color (the half-disc becomes whiter as standard deviation increases), and to variable linked to the size of the icon (the stroke at the border of the icon becomes wider as standard deviation increases).
4.1.5. Tilting the globe: showing changes in values through time

The interactive nature of three-dimensional globes such as WW makes it an ideal framework for visualizing the temporal aspects of data values. Depending on the location of the camera, the vertical dimension of the virtual earth can be made visible, showing the topography of the region of interest at the same time as opening a new space to display data. This extra dimension permits the visualization of data according to their “position” in time (MacEachren, 1995; Andrienko et al., 2003; Kraak, 2003). To represent changes in values through time, we developed the ability to flexibly alternate data views of variables having a temporal component. In the system described here, time values of variables sampled at a given site are shown when the user chooses to tilt the globe; alternatively, the mean and standard deviation is displayed when the point of view is directly vertical to the site. For each icon, an elevation index is dynamically assigned according to the date of sampling. For instance, if a variable has been sampled five times at a given site at a regular rate (for example, monthly), watching the scene from a tilted point of view will display each variable on a vertical time scale, where the oldest is represented closer to the ground, the most recent is shown higher towards outer space, and the three others are distributed in the space between them (Fig. 8a). A blank space is used if values are missing for certain dates, which leaves an empty slot. Alternatively, even if data have been sampled only once per site on a wide spatial range, tilting will quickly and effectively show indications of the time values of the samples. Selected variables are then not only shown as values, but as events in history. This global visualization of samples in their spatiotemporal context can potentially provide a better visual understanding of ecological patterns (Fig. 8b).
Fig. 8. a: Changes through time of the values of two variables on one site. The bottom icon is the oldest one, while the highest one is the most recently sampled. Users can easily see trends of the two variables in time. b: Two variables have been sampled only once on each site. As with Fig. 8a, the bottom icon is the oldest one, and the highest one is the most recently sampled. The difference in elevation quickly shows the chronology of the sampling process, and brings valuable information about a possible correlation between the values of displayed variables and their sampling date.

In total, these icons bear four dimensions (color, size, stroke thickness and half disc overlay transparency), which are to be added to the three spatiotemporal dimensions. The total of seven dimensions displayed at once does not match other spatial or aspatial multidimensional solutions (Chernoff, 1973; Kreuseler, 2000; Andrienko et al., 2003), but keeps visualizations simple enough to be understood at a glance. This, we believe, is an important criterion for a database geographic browsing tool such as this.
5. Discussion and annotation about datasets

Although essential, formal Metadata are rarely enough to describe all the details of original research, and informal communication is often very necessary (Zimmerman, 2007). There is theoretically no sufficient amount of metadata to implement in order to fully describe a dataset. The extra information needed for the effective second-hand use of data will therefore usually have to come from the authors themselves. When scientists consider reusing data sampled by other people in their own research, a few sensitive points arise (Zimmerman, 2003). Among them is the essential need for the person about to reuse data to have a sufficient level of understanding of the dataset (data acquisition methods used, condition of sampling, etc.), in order to build confidence in the compatibility of the original fieldwork with the new analysis. This confidence is mostly achieved by interpreting metadata in light of their own data acquisition experience. Moreover, in some cases metadata is not enough and personal communication with the authors of the original work is a necessary step (Zimmerman, 2007). To ease this communication, we implemented a module to discuss the datasets displayed on the virtual globe. For each visualization, a corresponding page showing a list of the featured datasets (with their corresponding geographic localizations) can be accessed. Users can then select one of the datasets to open a corresponding discussion board and ask questions or exchange constructive considerations about it with other researchers and with the author of the dataset.

6. How it all works together

The modular architecture of SFMN GeoSearch makes its use very flexible: all the modules described in this paper are linked together, and actions in one module lead to either results or new opportunities to explore further the data in another. Therefore, describing formal sequences of actions to extract and display data is difficult. Let us follow, to illustrate how the system can be interacted with, just one example of the use of the application for data discovery.

6.1. Displaying data on the virtual globe knowing the variables' names or site name

Once a variable has been selected in the tree menu (Fig. 1-c), its name is added to a list (Fig. 1-e) and the value of this variable is automatically shown on the virtual globe (Fig. 1-a) as an icon, everywhere it has been sampled. Also, the names of the authors of the studies from which the displayed values have been taken appear in the “Authors” list (Fig. 1-f). When variables have been sampled several times on the same site, the application automatically computes and shows their means, as well as an indication of their standard deviations. If carefully selected, the icons can, with their different colors and/or size, help users to understand trends through the landscape. Once again, as dates of sampling are important to comprehend changes of value through space, users can, at all times, incline their view of the globe to turn the spatial visualization into the spatiotemporal visualization described above. Similarly, users can retrieve data by knowing the name of the study site, which is accessible through the hierarchical tree menu. Sites can be accessed by either alphabetical ordered lists or by rough toponymic classification.

6.2. Displaying variables on the virtual globe knowing their authors' names

Users can, in the “Social Network” tab, explore the social connections within the research network and select authors of interest. Back at the main page, the list of authors will be filled with the selected names. No data is displayed at this point, and the names appear in grey, showing that they are not yet connected to displayed values. To start using these names for data visualization, and to select data related to a particular researcher, users need to double-click on a name in this list. The hierarchical tree menu then expands to display the work of the selected author. As previously explained, variables can then be added to the “Variable” list. This action will update the “Sites” list, while the name of
the author of the dataset will switch to white, showing that his/her data is actually shown, and the virtual globe will display the variable on the site studied by the researcher. Other researchers can be added by the same means.

6.3. Narrowing down a selection

All these queries can be narrowed down to a particular region; users can specifically choose, in the tree menu, which sites to display. Moreover, users can narrow down their selection to periods of time of interest by selecting series of dates in the timescale (Fig. 1-g).

7. Discussion

Today's environmental challenges often involve phenomena acting across spatial and temporal scales. Understanding the relationship of local patterns to broad-scale issues is an active area of research, and we believe that useful information can be derived from multi-disciplinary sources across potentially large areas. Working towards addressing this fundamental issue, the organization of science has progressively shifted towards larger-scale scientific collaborations, and led to the emergence of knowledge networks (Geuna et al., 2003). Networks such as LTER, which has been gathering research sites across the USA for almost 30 years, providing scientists with large amounts of spatially and temporally broad ecological data, are a good example of what has been accomplished in the last decades (Baker et al., 2000; Andelman et al., 2004). Comprehensive technologies to manage them (Jones, 2007), as well applications dedicated to exploring and visualizing the wealth of organized data made available, are taking shape. They use the Internet as a medium to connect users to the meta-database, and are most often distributed via web portals (Halpin et al., 2006; Flemons et al., 2007). SFMN GeoSearch builds on some aspects of these systems to provide novel shared data exploration and spatiotemporal visualizations.

Space is an important dimension in discovering patterns in ecological data, and most of the existing data sharing systems use some kind of cartographic plotting tools to display the results of user's queries. For the most part, these cartographic components are two-dimensional, which implies obvious limitations in terms of multi-dimensional data visualization. At least with regards to the addition of the third dimension, virtual globes are a promising alternative. While some projects already use external applications such as Google Earth to display the results of queries made via a web portal (Gemmell et al., 2007), few or none integrate three-dimensional cartography within the interface of a data sharing system. The interaction between the geographic component and the query page is then slowed down by the constant need for downloading and handling visualization files between the web portal and the visualization environment. One of the improvements provided by SFMN GeoSearch is the tight integration of a three-dimensional cartographic module within the query window. Users can seamlessly build queries and see their result instantly on the globe, within the same window. This allows users to try many combinations of queries quickly, making the exploration of potentially large databases much more efficient. In ecology, time is often an important component, and cartographic methods have been designed to display spatiotemporal events (either punctual or continuous) through the third spatial dimension (Hägerstrand et al., 1967; Wachowicz, 1999; Kraak, 2003). Beyond the realistic backdrop, virtual globes make space for displaying temporal information through their third dimension. We believe the sense of interaction, as well as the reactivity of the visualizations on the virtual globe allows for a better connection between the users and the database.
8. Future work

For the SFMN GeoSearch system to be useful to a wide audience of researchers and the public, data import, data permissions, and data export modules must be carefully developed. We have begun to develop simple data import tools for general use. These are designed to accept comma-delimited flat files, in which the first line represents column headers and subsequent lines represent data points. A flexible parser maintains a list of existing variables, asking the users at upload time whether each identified variable is already in use in the larger database. Data permissions can be set to reserve individual tables for a given data owner. Data export routines can produce comma-delimited output or, in some cases, KML files. Like data input, data output raises complex concerns: in particular, if temporal data were represented in a KML as an elevation (as in the tilted globe view), it cannot also be used to represent a real-world elevation value (which may be an attribute of the sampled data). Recent advances in the KML specification to encode time values may make KML representation easier, but there are likely to be ongoing ontological complications on both the input and output side. There are additional ways to substantially improve the existing aspects of SFMN GeoSearch. These include: (1) Tools and approaches to manage database consistency. Gathering ecological variables collected by different researchers for different goals, with various methods and equipment, leads to heterogeneity in the database. Likewise, similar variables are often named through different terms. Ontology and metadata management can begin to address these essential problems. These are active domains of research providing new solutions for formal description and effective management of ecological data. We believe that the integration of metadata management systems such as Metacat (Jones et al., 2001) could greatly improve the overall quality of SFMN GeoSearch's database; (2) Improved geographic exploration. Users should be able to use the virtual globe as much for visualizing as for building queries. The exploration of the database will therefore be pushed towards a deeper integration of the geographic tool as an exploration tool. For instance, drawing a rectangle on the globe, or simply zooming toward a smaller portion of territory, would narrow the query to the specified spatial extent; (3) Data visualizations. For now, our system is designed to display punctual quantitative values. An important next step would be to implement ways to display qualitative data, as well as integrating additional means to visualize scientific data through, for example, raster maps, polygons or lines; and (4) User design and feedback. We need to polish the user interface and improve the general user experience by collaborating with a sample of researchers from the network. Although this system has been designed with point-based forestry data foremost in mind, we believe the approach generalizes to a wide array of point-based data sources. We imagine several other applications of the framework, including water quality data, air sampling values, weather-related observations, or illustration of data from GPS-equipped animals. With minor enhancements, it should be straightforward to support links to real-time webcam data or other data provided by a link to a frequently updating KML file.

9. Conclusion

SFMN GeoSearch is designed to help members of scientific networks to share, discover, archive, and combine ecological field data. In the long run, system modules are expected to receive georeferenced datasets of field or modeled ecological data, to organize them, to allow users to explore and to visualize them intuitively, and to download relevant datasets. Currently, the data visualization modules are the most mature, and data can be explored through: (1) a hierarchical menu giving access to the whole database, or (2) a network visualization interface that provides clear and fast ways to discover data through their author's social connections. The latter can efficiently address problems related to the lack of social connection awareness within the social network. Our system also helps facilitate discussions between members of the network (and hopefully between the authors and the potential
secondary users) about datasets contained in the database. Although our system has been built for the SFM Network, it is an open platform and can be adapted to most other ecological research networks.

Ecological variables are tightly connected to their spatial and temporal contexts (Kreuseler, 2000), and we believe visualizing aggregated datasets along these spatiotemporal dimensions is important for data discovery in large metadata databases. When displayed at the same time, ecological data produced and managed by different people at different dates for research relative to different sites can then be understood: (1) within their geographical and temporal contexts, (2) across different scales, since the change in a variable's values through time can be visualized just as easily at the continental or site level, and (3) in comparison to each other. A tool like SFMN GeoSearch, leveraging novel and interactive database exploration and spatiotemporal visualizations, can facilitate our understanding of trends through space and time in such a large amount of shared data, effectively assisting users in spotting the right datasets to reuse in new research.

10. Extra material

A demo of SFMN GeoSearch can be found at: http://meta.geog.umontreal.ca/currentprojects/forests/SFMN/Welcome.html. We have developed a substantial amount of information at that site, including descriptions, videos, and demo versions of the data browser, the network browser, and another related sub-project. We will keep these links active for the foreseeable future. Videos at the site require no special software; demo versions require a recent personal computer (running either Microsoft Windows, Max OS X, or Linux/Unix), with several gigabytes of memory and a 3D accelerated graphic card. On the software side, an up-to-date web browser such as Mozilla Firefox, and a recent Java runtime environment are needed (the latter two are free and widely available).

Acknowledgments

We thank The Sustainable Forest Management Network and the Centre d'Étude de la Forêt (Forest Study Centre) for their financial support. We thank the SFMN proposal evaluation board for approval of this project and Dr. Jim Fyles for providing regular feedback and support during the development of the SFMN GeoSearch system.

References


