Towards Supporting Collaborative Data Analysis and Visualization in a Coastal Margin Observatory

Emanuele Santos, Phillip Mates, Erik Anderson, Brad Grimm, Juliana Freire, Cláudio Silva

Scientific Computing and Imaging Institute, University of Utah
Salt Lake City, UT
{emanuele, mates, eranders, bgrimm, juliana, csilva}@sci.utah.edu

ABSTRACT

Managing and understanding the large amounts of scientific data is undoubtedly one of the most difficult research challenges scientists are facing today. As large interdisciplinary groups work together, the ability to generate a diversified collection of analyses for a broad audience in an ad-hoc manner is essential for supporting effective scientific data exploration. Science portals and visualization web sites have been used to simplify this task by aggregating data from different sources and by providing a set of pre-designed analyses and visualizations. However, such portals are often built manually, and are not flexible enough to support the vast heterogeneity of data sources, analysis techniques, data products, and user communities that need to access this data. In this paper we describe a system that adopts the model used by social Web sites and that combines a set of usable tools and a scalable infrastructure for users to explore and re-use visualization and analysis pipelines. We describe our efforts on implementing such a site for the NSF Science and Technology Center for Coastal Margin Observation & Prediction (CMOP).

Author Keywords

Web collaboration, provenance, scientific workflows

ACM Classification Keywords

H.5.3 Information Interfaces and Presentation: Group and Organization Interfaces—*Collaborative computing*

INTRODUCTION

One of the most difficult research challenges that scientists face today is managing and understanding the enormous amounts of scientific data that are generated. As large interdisciplinary groups need to collaborate, the ability to generate a diversified collection of analyses for a broad audience in an ad-hoc manner is essential for supporting effective exploration of scientific data.

The NSF Science and Technology Center for Coastal Margin Observation & Prediction (CMOP) [2] is a multi-institutional center dedicated to coastal margins, which are regions consisting of very productive ecosystems, susceptible to different scales of variability, and that play an important role in global elemental cycles. CMOP maintains the Science and Technology University Research Network (SATURN) observatory: a network of heterogeneous observation platforms coupled with large-scale simulation models of ocean circulation. The platforms consist of fixed and mobile stations with different sensors measuring physical properties, such as temperature, salinity and water level; and biochemical properties, such as nitrate, chlorophyll and dissolved organic matter concentrations. Each of these sensors may generate over a million measurements in a couple of days. Simulation results are generated by two systems: a suite of daily forecasts targeting specific estuaries, and long-term hindcast databases, where the simulations are re-executed using observed data as inputs. All this data together is used to predict oceanographic features with practical realism.

Because of the broad influence of coastal margins, there is an intrinsic heterogeneity in data sources, analysis techniques, data products, and user communities, which makes it challenging to design a system flexible enough to be used by scientists, policy makers, students, and the general public. Besides, in interdisciplinary environments like CMOP, there is a considerable technological learning curve for scientists to use specialized libraries for manipulating data and deriving data products. Even for experienced users, there are no accepted "best practices" that ensure the wealth of information produced by observations, predictions and analysis is effectively used.

Science portals [1, 10, 6] and visualization websites [15, 14] have been used to simplify data exploration by aggregating data from different web sites and by providing a set of canned analysis and visualization tools. However, they are insufficient for handling large volumes of heterogeneous data and the diversity of stakeholders and their needs. It is simply not possible for IT personnel to anticipate all necessary analyses and different ways to correlate and integrate data. Furthermore, while some analyses that are used regularly can be canned, others are ground-breaking and need to be created, altered on-the-fly, and improved as part of a collaborative effort. At CMOP, currently, on-the-fly generation

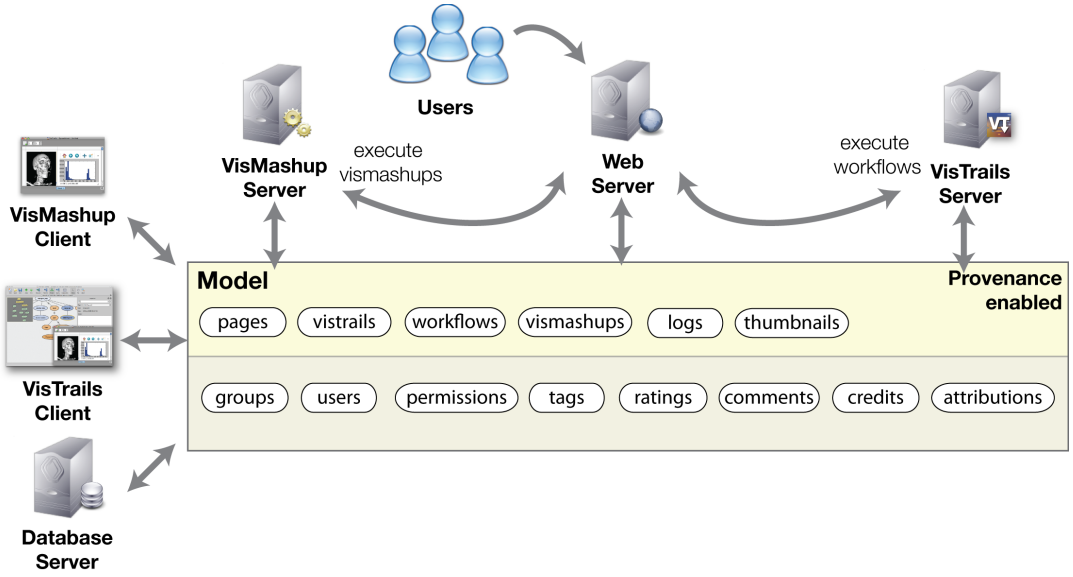

Figure 1. Architecture of the social data analysis site based on VisTrails and VisMashup. The VisTrails provenance model is shared with all the other other servers and client applications in the architecture. The data model is also extended to include the social Web features such as groups, permissions, ratings and comments. Mashups and workflows embedded in web pages are executed in the servers and the results are displayed in the web page.

of visualization data products is often avoided, since creating new analyses and publishing their results is time-consuming, often requiring programming expertise and a trial-and-error cycle, demanding intense and off-hours interactions of IT staff with scientists.

In order to enable scientists to effectively perform their own analyses, and to create and publish new data products, we need to simplify authorship and customization of these analysis and visualization tools, and try to minimize the need for the intervention of expert programmers and IT personnel.

In this work-in-progress paper, we introduce our attempt to build a system that enables the construction of transformative applications and data products by interested parties with diverse goals in science, education, and public policy. To accomplish this task, we adopt the model used by social Web sites [8, 3] and Web-based communities [4, 16] and develop tools to enable social analysis of scientific data. Our goal is to facilitate collaboration and sharing among users, not only of data but also of analyses. Shared repositories of analysis and visualization workflows will expose users to a large number of tasks that provide examples of (sophisticated) uses of tools. By querying the workflow specifications, along with data products and their provenance, users can leverage the collective wisdom to learn by example from the reasoning and/or analysis strategies of experts; expedite their scientific training in disciplinary and inter-disciplinary settings; and potentially reduce the time lag between data acquisition and scientific insight. Shneiderman [13] points out that "much of our intelligence and creativity results from interactions with tools and artifacts and from collaborating with other individuals". To this end, we combine a set of usable tools and a scalable infrastructure for users to explore and re-use visualization and analysis pipelines. We describe an initial implementation applied at CMOP.

SUPPORTING A SCIENCE COLLABORATORY

To allow scientists to share, re-use and collaboratively design computations, monitoring tasks, and analyses, our infrastructure is based on scientific workflows. Our system builds on VisTrails [12] and VisMashup [11], but adds a number of other components. The architecture of our social data analysis site using the infrastructure described above is illustrated in Figure 1. The collaboratory infrastructure shares the provenance model with the other servers and client applications in the architecture. The model also incorporates the social web features, such as comments and ratings and can also be serialized to a database server.

It builds on VisTrails for its workflow functionality. In our existing implementation, users need to use VisTrails on their local machine to create workflows that are uploaded to a workflow server. As part of the VisTrails project, we have developed several basic components and intuitive interfaces that support many of the tasks required to support a science collaboratory, including: visual difference interface that allows structural comparison of workflows [5]; a query-by-example interface that allows users to quickly construct expressive queries using the same familiar interface they use to build workflows [12]; a mechanism whereby users can refine workflows by analogy, i.e., users to can perform complex modifications to workflows without requiring them to directly modify the workflow specifications [12]; a system that mines workflow collections and learns common paths which are then used to derive recommendations during workflow design process, suggesting potential modules and connections in a manner similar to a Web browser suggesting URLs [7].

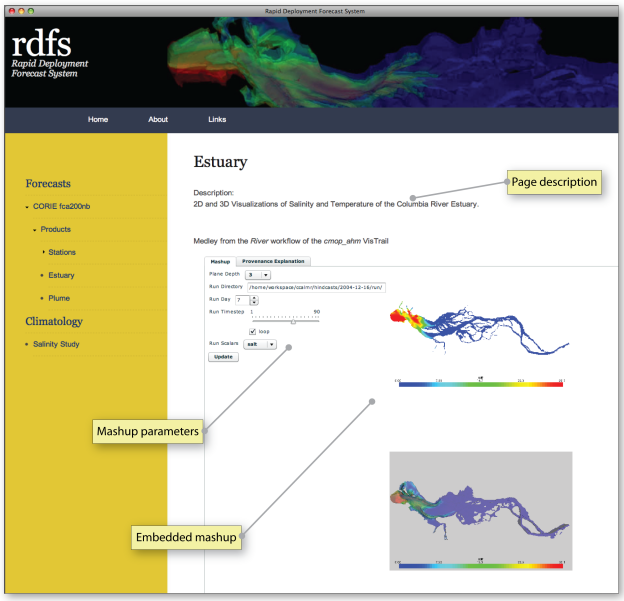

Figure 2. Interacting with a mashup in the RDFS web site. Users can see 2D and 3D visualizations of salinity or temperature in the Columbia River Estuary at different depths.

The visualizations are built using VisMashup, which is a workflow-based framework for creating mashups and simplify the process of publishing (and sharing) scientific results [11]. After the workflows are uploaded to the database, workflow designers can use the VisMashup framework to create customized applications or mashups based on these workflows. Because these mashups can be customized for very specific tasks, they can hide much of the complexity in an data analysis or visualization specification and make it easier for users to explore visualizations by manipulating a small set of parameters. VisMashup automatically generates a graphical user interface based on the select parameters that can be deployed either on the desktop or on the Web. Authenticated scientists on the web site can choose from a collection of mashups and add them to the web pages. If no interaction is necessary, only the results of workflows can also be added to the web pages. In order to execute mashups and workflows, the web server communicates with a server version of VisTrails and VisMashup. Ideally, these servers can be deployed on a cluster or other high performance architectures. This allows the use of very large datasets, which is not possible by current tools. The system also caches executions, so images requested by already executed mashups are reused.

Through the mashups, users can also access the *provenance explanation* of the mashup, *i.e.,* the system uses the provenance information to show how that data product was generated. In contrast to Many Eyes [15] and its Wikified version [9], our system is not dataset-oriented. Workflows and mashups are the basic sharable objects. Users with the proper permissions can reuse, rate and comment on workflows and mashups.

RAPID DEPLOYMENT FORECAST SYSTEM (RDFS)

We have used our system for reimplementing some existing CMOP functionality in the form of the Rapid Deployment Forecast System (RDFS), see Figure 2.

The Rapid Deployment Forecast System (RDFS) is designed to facilitate the creation and implementation of new forecasts. While developing the system, visualization tools were created that continue to be used by many of CMOP modelers – especially for creating GIF animations of the forecast data. In the way RDFS was initially designed, only modelers were able to create new forecasts. However, by using our system, mashups and workflows representing the visualization tools were added to the database and users could easily reuse them for different simulation forecast models. More flexible visualizations that allow parameter changes on-the-fly were generated. Users can also choose to execute the mashups in their local machines so as to make complete use of 3D navigation. Different types of mashups according to the audience can be used on the system, so scientists interested in more details about the salinity values at a specific station can use a mashup similar to the one in Figure 3(a). The provenance explanation for this mashup is illustrated in Figure 3(b).

DISCUSSION

We are still in the process of deploying the new revamped RDFS system based on our architecture. Initial feedback from CMOP scientists and IT staff has been positive, but there are a number of feature requests that we are currently implemeting.

In our current system, some operations still need the installation of VisTrails on local machines, in particular, for the creation of new workflows. This will continue to be the case while we do not fully finish the implemementation of a Web interface for the VisTrails Builder that we have been pursuing for the past few months.

A more serious issue is how to properly support better interaction with high-end 2-D and 3-D data products over the web. Since our workflows support a large collection of underlying libraries, such as VTK, Matplotlib, ImageMagick, it is not really feasible to completely reimplement them for use over the web. One possibility would be to augment VisTrails Spreadsheet cells with improved web support, and interactivity, although not to the point of supporting everything those libraries might support natively.

CONCLUSIONS

We introduced a Web-based system that enables the construction of transformative applications and data products by a broad audience. To accomplish this task, we adopted the model used by social Web sites and Web-based communities and developed tools to enable "social analysis of scientific data". Our goal is to facilitate collaboration and sharing among users, not only of data but also of analyses, by combining a set of usable tools and a scalable infrastructure for users to explore and re-use visualization and analysis pipelines. We also described our efforts on implementing an

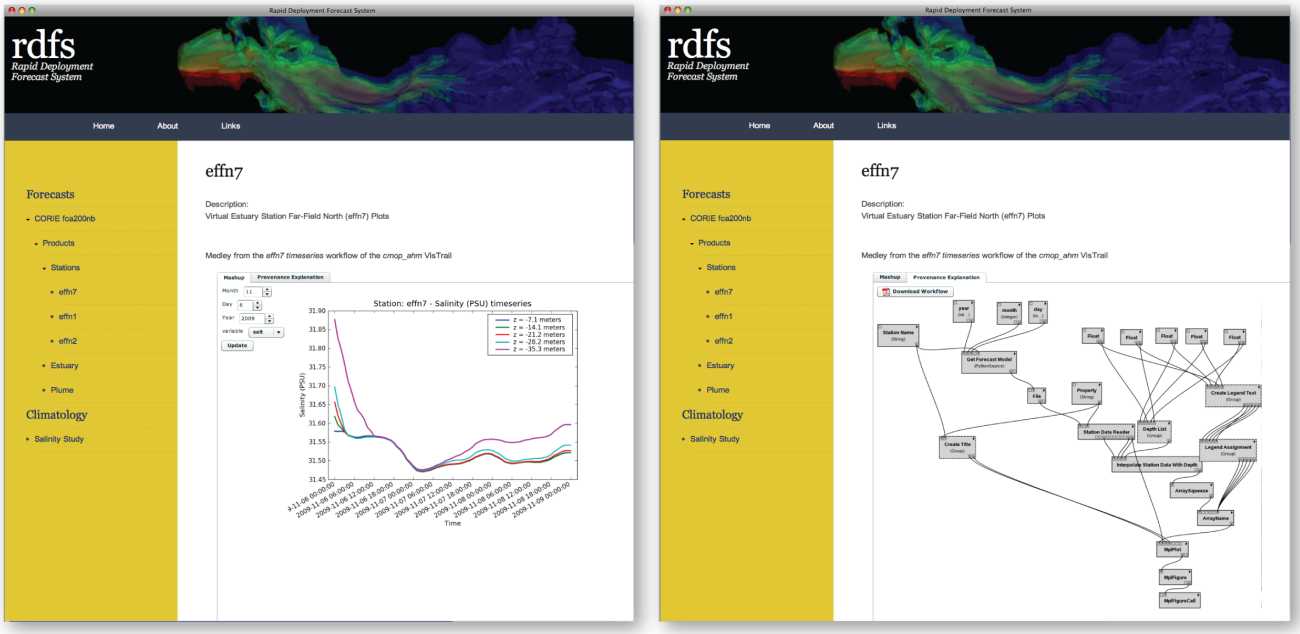

Figure 3. (a) Mashup displaying plots of salinity and temperature measurements at different depths for a specific station and (b) its provenance explanation.

initial version of the system and applying it in the context of an ocean observatory.

ACKNOWLEDGMENTS

This work is partially supported by the NSF (under grants IIS-0844572, CNS-0751152, IIS-0746500, IIS-0513692, CCF-0401498, EIA-0323604, CNS-0514485, IIS-0534628, CNS-0528201, OISE-0405402), the DOE, and an IBM Faculty Award. E. Santos is partially supported by a CAPES/Fulbright fellowship.

REFERENCES

1. Chemical blogspace. http://cb.openmolecules.net/.
2. NSF Center for Coastal Margin Observation and Prediction (CMOP). http://www.stccmop.org.
3. Facebook. http://www.facebook.com.
4. Flickr. http://www.flickr.com.
5. J. Freire, C. Silva, S. Callahan, E. Santos, C. Scheidegger, and H. Vo. Managing rapidly-evolving scientific workflows. In *International Provenance and Annotation Workshop (IPAW)*, LNCS 4145, pages 10–18. Springer Verlag, 2006.
6. R. Hoffmann. A wiki for the life sciences where authorship matters. *Nature Genetics*, 40(9):1047–1051, 2008.
7. D. Koop, C. Scheidegger, S. Callahan, J. Freire, and C. Silva. Viscomplete: Data-driven suggestions for visualization systems. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1691–1698, 2008.
8. Many eyes. http://services.alphaworks.ibm.com/manyeyes/home.
9. M. McKeon. Harnessing the Web Information Ecosystem with Wiki-based Visualization Dashboards. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1081–1088, 2009.
10. A. R. Pico, T. Kelder, M. P. van Iersel, K. Hanspers, B. R. Conklin, and C. Evelo. WikiPathways: Pathway editing for the people. *PLoS Biology*, 6(7), 2008.
11. E. Santos, L. Lins, J. Ahrens, J. Freire, and C. Silva. Vismashup: Streamlining the creation of custom visualization applications. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1539–1546, 2009.
12. C. Scheidegger, D. Koop, H. Vo, J. Freire, and C. Silva. Querying and creating visualizations by analogy. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1560–1567, 2007.
13. B. Shneiderman. Creativity support tools: accelerating discovery and innovation. *Commun. ACM*, 50(12):20–32, 2007.
14. Swivel. http://www.swivel.com.
15. F. B. Viegas, M. Wattenberg, F. van Ham, J. Kriss, and M. McKeon. ManyEyes: A site for visualization at internet scale. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1121–1128, 2007.
16. YouTube. http://youtube.com.