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 volumes of scientific data is one of the most difficult challenges scientists face today. As interdisciplinary groups work together, the ability to generate a diversified collection of analyses for a broad audience and in an ad-hoc manner is essential to support effective data exploration. Science portals and Web-based visualization tools have been used to simplify this task by aggregating data from different sources and providing a set of pre-defined analyses and visualizations. These, however, are expensive to build and lack the flexibility necessary to support the vast heterogeneity of data sources, analysis techniques, and information needs from multiple user communities. In this paper, we present a system that adopts the model used by social Web sites and, by combining a set of usable tools and a scalable infrastructure, simplifies the construction of science collaboratories: Web sites where groups of users can collaboratively explore scientific data. An important feature of the system is that it allows users to easily customize and publish new analyses and visualizations on the Web. We also describe our implementation of a collaborative site for the NSF Science and Technology Center for Coastal Margin Observation & Prediction (CMOP).

Author Keywords

Web collaboration, provenance, scientific workflows,mashups

ACM Classification Keywords

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

INTRODUCTION

As 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. Consider the requirements of the NSF Science and Technology Center for Coastal Margin Observation & Prediction (CMOP) [2]. CMOP 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.

Like CMOP, several other scientific projects struggle to provide the necessary infrastructure to enable users to effectively analyze their data. A popular solution has been the creation of science portals [1, 11, 7]. These portals, however, are costly to create and maintain and the analyses they provide support only small subset of the data products users need. It is simply not possible for IT personnel to anticipate all necessary analyses and different ways to correlate and integrate data. And 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, on-the-fly generation of visualization data products is often avoided, since creating

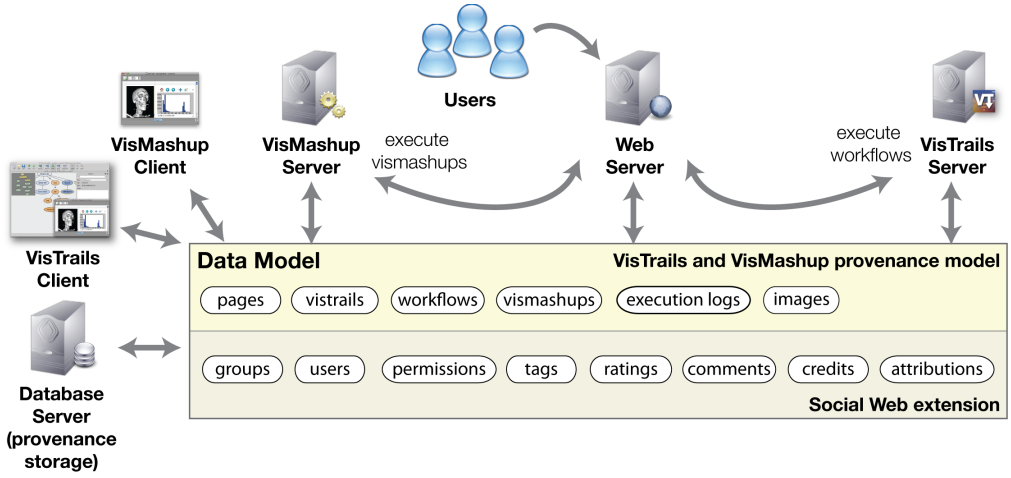

Figure 1. Architecture of the social data analysis site based on VisTrails and VisMashup. The VisTrails provenance model is shared with all the other servers and client applications in the architecture. The data model is also extended to include the social Web features such as groups, permissions, ratings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

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.

More recently, Web-based visualization tools [16, 15] have been used to simplify data exploration by aggregating data from different Web sites and by providing a set of canned visualization techniques. But these are not scalable: they neither are able to handle large volumes of heterogeneous data, nor the diversity of visualization and analysis techniques required by stakeholders.

In this paper, we describe our preliminary work on addressing these limitations: a framework that simplifies the construction of science collaboratories. We adopt the model used by social Web sites [9, 4] and Web-based communities [5, 17] to develop tools that 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 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 [14] 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 propose a system that combines a set of usable tools and a scalable infrastructure for users to explore and re-use visualization and analysis pipelines. Besides enabling scientists to perform their own analyses, the system also simplifies the creation and publication new data products, reducing the need for interactions with IT personnel. We describe an initial implementation of the system for CMOP.

BUILDING SCIENCE COLLABORATORIES

Our system allows scientists to share, re-use and collaboratively design computations, monitoring tasks, and analyses that are specified as scientific workflows [3]. It uses the infrastructure provided by VisTrails [13] and VisMashup [12], and adds a number of new components. The high-level architecture of the system is illustrated in Figure 1. The current VisTrails and VisMashup provenance data model is shared with all client and server applications in the system. The model was extended with social Web elements, including groups, users, ratings and comments. Mashups and workflows embedded in Web pages are executed in the servers and the results are displayed on the same pages. Below we describe how the system is used to provide a rich collaborative environment.

Designing Workflows. Users (scientists or developers) use VisTrails on their desktops to create workflows that are uploaded to a workflow database. VisTrails provides several tools 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 [6]; a query-by-example interface that allows users to quickly construct expressive queries over a workflow collection using the same familiar interface they use to build workflows [13]; 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 [13]; 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 [8].

Another important feature of VisTrails we use is the detailed provenance it captures both of the workflow design process and execution [6]. As we discuss below, this information is essential for the creation of workflow mashups.

Creating Workflow Mashups. After the workflows are uploaded to the database, workflow designers can use the VisMashup framework [12] to create customized applications or mashups based on these workflows. By allowing the creation of simplified views of workflows, VisMashup simplifies the process of publishing (and sharing) scientific results. Because these mashups can be customized for very specific tasks, they can hide much of the complexity in a data analysis or visualization specification and make it easier for users to explore data products 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.

Publishing Customized Analysis and Visualizations. Authenticated scientists on the Web site can choose from a collection of mashups and add them to the Web pages. Once these pages are saved, users can interact with the mashups from their browsers (see Figure 2). When a user inputs a set of parameters in the interface, the underlying workflow is executed. Note that if no interaction is necessary, only the results of workflows are added to the Web pages. In order to execute mashups and the underlying workflows, the Web server communicates with a server version of VisTrails and VisMashup. Ideally, these servers should 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 [16] and its Wikified version [10], 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 to implement a new version of the CMOP Rapid Deployment Forecast System (RDFS) (Figure 2). RDFS was designed to facilitate the creation and implementation of new forecasts. It includes several visualization tools created for CMOP modelers that are used to generate GIF animations of the forecast data. The first version of the system was used exclusively by modelers to create new forecasts.

In the new version created using our system, mashups and workflows representing the visualization tools were added to the database. Consequently, users can easily access and reuse them for different simulation forecast models. In addition, more flexible visualizations that allow parameter changes on-the-fly were generated. While users can interact with mashups over a Web browser, they can also choose to do so in their desktop so that they can make full use of 3D navigation.

Different types of mashups can be easily created depending on the information needs of a user (or group of users). For example, 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).

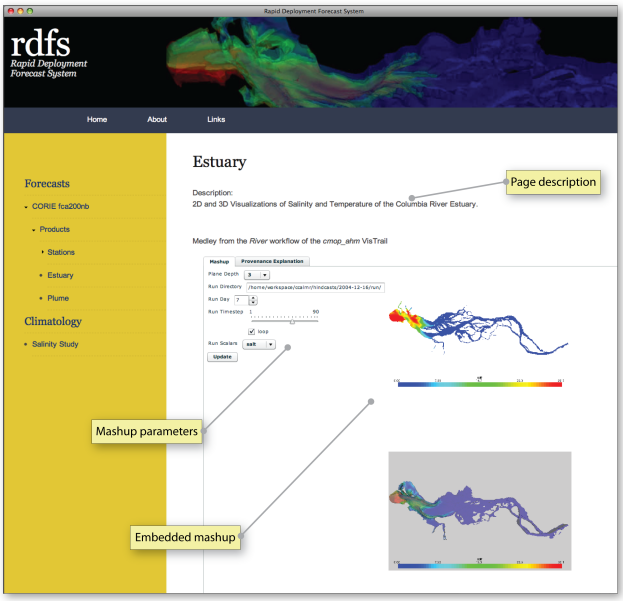

Figure 2. Interacting with a mashup in the RDFS Web site. Users can see 2D and 3D visualizations of salinity or temperature in the Columbia River Estuary at different depths.

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 implementing.

An important open problem we have been investigating is how to 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 re-implement 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 have introduced a system that simplifies the creation of science collaboratories by enabling the construction and publication of customized applications and data products on the Web. 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". The system facilitates 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 initial version of the system and applying it in the context of an ocean observatory.

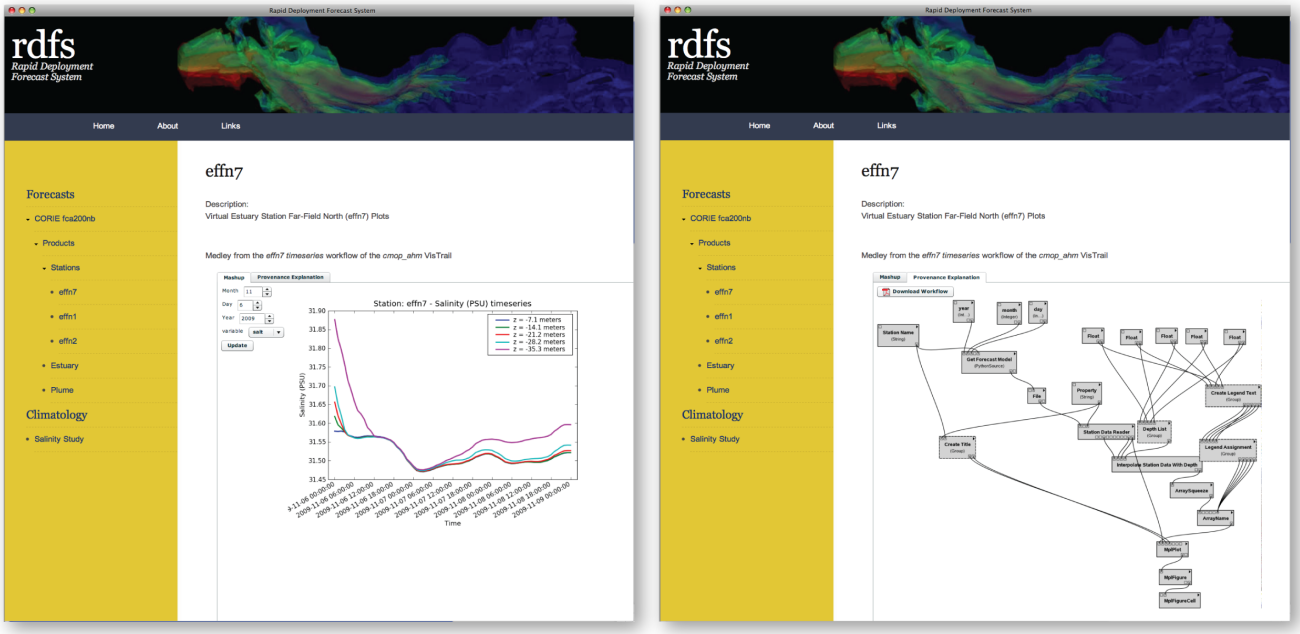

Figure 3. (a) Mashup displaying plots of salinity and temperature measurements at different depths for a specific station and (b) its provenance explanation.

ACKNOWLEDGMENTS

This work is partially supported by the NSF (under grants IIS 0905385, IIS-0844572, CNS-0751152, IIS-0746500, IIS-0513692, CNS-0528201, OCE-0424602, OISE-0405402), the DOE, and IBM Faculty Awards. 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. S. B. Davidson and J. Freire. Provenance and scientific workflows: challenges and opportunities. In *ACM SIGMOD*, pages 1345–1350, 2008.
4. Facebook. http://www.facebook.com.
5. Flickr. http://www.flickr.com.
6. 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.
7. R. Hoffmann. A wiki for the life sciences where authorship matters. *Nature Genetics*, 40(9):1047–1051, 2008.
8. 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.
9. Many eyes. http://services.alphaworks.ibm.com/manyeyes/home.
10. M. McKeon. Harnessing the Web Information Ecosystem with Wiki-based Visualization Dashboards. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1081–1088, 2009.
11. 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.
12. 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.
13. 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.
14. B. Shneiderman. Creativity support tools: accelerating discovery and innovation. *Commun. ACM*, 50(12):20–32, 2007.
15. Swivel. http://www.swivel.com.
16. 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.
17. YouTube. http://youtube.com.