Leveraging Web 2.0 technologies in a Cyberenvironment for Observatory-centric Environmental Research

Yong Liu, Jim Myers, Barbara Minsker, Joe Futrelle

National Center for Supercomputing Applications
1205 W. Clark St.
Urbana, IL 61802, USA

{yongliu, jimmyers, minsker, futrelle}@ncsa.uiuc.edu

ABSTRACT

In this paper, we will highlight how Web 2.0 technologies and patterns have been successfully applied in one of the National Center for Supercomputing Applications (NCSA) projects, which is supporting environmental research and the developing U.S. WATERS (**WAT**er and **E**nvironmental **R**esearch **S**ystems, an initiative supported by US NSF GEO and ENG directories) network observatories. Driven by real scientific use cases arising from the current active investigators involved in testbed projects for the WATERS network, these highlights include user-generated content, rich user experience with AJAX and innovative usage of mashup with Google Maps, contextualized scientific knowledge network creation with extensible, folksonomic provenance support, and events propagation with RDF-enhanced messages exchange. The resulting system provides state-of-the-art collaboration capabilities and is both far more flexible in its ability to interact with desktop applications and community services and more amenable to third-party customization than early systems. In addition to describing these features in more detail, we will also discuss future directions that continue Web 2.0-style thinking towards the semantic grid, i.e. through the use of lightweight global identifiers and a content management approach to metadata that enables use of multiple ontologies and dynamic evolution (e.g. folksonomies) of terminology.

URL: http://cleaner.ncsa.uiuc.edu/cybercollab

Keywords

Environmental Observatory, Web 2.0, User-Generated Content (UGC), Mashup, AJAX, Resource Description Framework (RDF), Scientific Knowledge Network, Cyberenvironment, Provenance, workflow.

1. INTRODUCTION

Current and coming initiatives for building large-scale environmental observatories, such as the **WAT**er and **E**nvironmental **R**esearch **S**ystems (WATERS) network, Ocean Research Interactive Observatory Networks (ORION), National Ecological Observatory Network (NEON), etc., for scientific and societal usages, see a need for an integrated cyberenvironment to support decentralized research. *Indeed, the ability to provide community-scale infrastructure while enabling innovation by individual researchers is a central challenge for cyberinfrastructure/e-science efforts.* The National Center for Supercomputing Applications (NCSA) has initiated efforts in building end-to-end cyberenvironments that provide this flexibility [1]. Since September 2005, the **E**nvironmental **C**yber**I**nfrastructure **D**emonstrator (ECID) Project [1] has successfully leveraged Web 2.0 technologies and patterns to build an integrated cyberenvironment to support real-world use cases of environmental observatories.

***OGF-19 Semantic Web 2.0 and Grid Workshop,* January 29, 2007, Chapel Hill, NC, USA.**

2. ENVIRONMENTAL OBSERVATORIES USE CASE

The WATERS planning effort is motivated by the national need to understand and restore lake, stream, and coastal water quality to achieve sustainable and secure water supply while improving and preserving aquatic habitats [2]. A proposed tiered/multi-scale remote and embedded sensing network will enable researchers to answer critical questions with regard to spatiotemporally-distributed hydrologic and environmental phenomena in a timely fashion. Methods are needed to access and control multiple data sources in real time, enabling adaptive monitoring and integrated management of water resources across water-monitoring infrastructures. Adaptive models are needed that use feedback from observation to improve the ability to simulate and predict behavior.

A challenging use case from one of the WATERS test projects in Corpus Christi Bay of Texas [3] requires support for the full lifecycle of scientific research. These researchers are working to apply sophisticated models to streaming sensor data to identify sensor anomalies and to forecast conditions such as low dissolved oxygen (also known as “hypoxia”). This requires the ability for researchers to apply models built as workflows to the data streams and to publish their derived results as new streams available to the community.

3. LEVERAGING WEB 2.0 TECHNOLOGIES AND PATTERNS

In this section we highlight the Web 2.0 technologies and patterns that have been used in the development of cyberenvironments for environmental observatory research.

3.1 User-Generated Contents (UGC)

User-Generated Content (UGC) is the focus of TIME magazine's 2006 Person of the Year award, which went to all Internet users who contribute to user-generated media such as YouTube and Wikipedia. Likewise, user-generated content and knowledge play a large role in the environmental observatory system. In our current system we have incorporated a component that enables community users to write individual blogs or post announcements

to group blogs. We also have integrated a mediaWiki-based wiki system for supporting collaborative writing, which the WATERS project planning committees have been actively using. Discussion boards (or forums) are also being used and integrated with mailing lists so that bi-directional posting (email to forum and forum to email) is supported. Other documents such as Word or PDF file or PowerPoint slides are kept in document libraries where users can also post threaded comments on individual files.

In addition, we think that user-generated contents should also include investigator-contributed workflow templates, models, and algorithms. Thus we allow users to publish their workflow templates and share them with the entire community who have access to the observatory resources. Such workflow publishing/sharing concept can also be found at myExperiment [4], an initiative in UK's myGrid research community. Furthermore, as demonstrated at SC06 [5] and AGU Fall 2006 meetings [6][7], users are allowed to change the workflow parameters on-the-fly and publish the new workflow to a server which can then produce new data streams for community use--this goes beyond just publishing workflow templates for others to use locally, but truly lets users start to add new resources to an observatory system. This is central to end-user customization and can truly facilitate individual innovation and community collaboration. Users of such integrated cyberenvironments for environmental observatories are no longer just passive consumers (e.g., getting data from the observatory), but also active participants and contributors. This resembles users' roles in a Web 2.0 environment.

All these user-generated contents can be and have been harvested through a daily log-style provenance service, as described further in section 4.

3.2 AJAX and Mashup with Google Maps

Environmental data are inherently geo-referenced, and such geo-contextual information should be preserved for a rich user experience. AJAX provides much more interactive navigation of web browser-based applications, while mashup provides extensible capability using a third-party published API to integrate external resources and applications. In our project we integrated Google Map APIs with user subscriptions service where users can find and subscribe to both raw data streams and derived data streams (such as a sensor anomaly data stream generated from an anomaly detection workflow running continuously in the observatory server) from the sensor platforms identified on Google Map. Pushlets, AJAX and Spring 2.0 asynchronous message driven POJO (Plain Old Java Object) have also been integrated to automatically plot/update the real-time sensor data streams (both raw and derived) inside the web browser [6]. Additional extra layers of information about the sensor, sensor platform,s and surrounding geographical features, either obtained from observatory GIS (Geographical Information Systems) servers or human-observed user-captured digital images shot by field scientists using cellphone/digital cameras about the sensors/measurements or other relevant environmental phenomenon, could also be overlaid on the sensor map.

Such mashup capability promotes end-user innovation since users can now share more resources and access a more seamless integrated virtual observatory. In additional, the augmentations of external resources and applications add new provenance information that can enrich semantic traceability. Specifically, a geo-referenced provenance trail would illustrate how sensor data in a particular geophysical location is used in certain models, workflows, and publications by certain people. This will help observatory users to evaluate the quality of the data and build trust among the community.

3.3 Contextualized Scientific Knowledge Network

One of the noticeable patterns of Web 2.0 is the social network capability provided by web sites such as MySpace and Facebook. Such capability has been leveraged in our project with much more extended and contextualized scientific information [8]. Users can browse contextualized scientific knowledge network in a clickable graphical network format, where documents, publications, people, workflow, data, etc. related to the user's task at hand are readily accessible and recommended. Such network creation was supported by the ubiquitous provenance semantic service, which we will describe in detail in section 4.

3.4 Event Propagation using RDF Enhanced Messages

Commonly seen in large-scale heterogeneous grid computing environments, message-based integration technologies are well-accepted in the web industry and, more gradually, in scientific communities. To enhance the semantic information exchange between different event generators and consumers, we embedded RDF triples in the JMS (Java Messaging Service) messages, which help to prevent the loss of critical information when the event propagates in the software components' consumer chains. This is significant in that a community-scale cyberenvironment usually loosely couples lots of heterogeneous software components. Augmenting machine-understandable RDF triples in the message facilitates such community-wide integration.

4. UBIQUITOUS PROVENANCE SERVICE

As indicated in the previous section, a provenance service has been used to integrate heterogeneous resources. The current approach taken in this project is to enforce global unique identifier for identities that are significant for integration, and our initial implementation has focused on agreement on unique global user identifiers across different components [9]. The underlying Tupelo 2 system developed at NCSA provides log-style API for different components (e.g., wiki, discussion board, document library, or workflow) to record RDF triples. The resulting RDF triples are stored separately on triple stores (currently KOWARI). The separation of the actual data store and the metadata store provides decoupling between the provenance service and other regular content or system components, resulting in the ubiquitous collection of provenance information.

While individual applications/software components may have ontologies (implicit or explicit, but a given tool only generates a particular limited set of triples), the triple store is RDF-based and accepts any valid triples from any source. Using common identifiers in that system then allows the capture of the overall network of relationships and the social network/recommender style features described in previous section.

5. CYBERENVIRONMENT MASHUP, FOLKSONOMIES AND MULTPLE ONTOLOGIES

The continuing development of an integrated cyberenvironment for environmental observatories relies on further development of the middleware that supports services where users can readily use published cyberenvironment APIs or Web Services to do "mashup" style integration either in their own familiar software environments (e.g., Matlab) or in an already-integrated cyberenvironment such as the one recently built by the ECID project.

Thanks to RDF's incremental, "bottom-up" style support for metadata integration, we will continue investigating using global unique identifiers for identities that are significant for integration. For example, because of the geo-referenced nature of most the data streams from environmental observatories, the spatial location can be a very strong integration point where location-based data service can be provided for downstream data consumers. This service should not just restrict to point source, but also support spatial query (e.g., temperature measurements within a radius of 5 miles).

In addition, multiple ontologies will be continuously incorporated. Dublin-core and other folksonomy-style ontologies have already been used. Other ontologies such as the CUAHSI Observation Data Model (ODM) [10] will be incorporated when they are finalized and used in the observatory system.

6. CONCLUDING REMARKS

We have leveraged several important Web 2.0 technologies/patterns in our ECID project development and have demonstrated far more flexibility in terms of providing collaboration, coordination and community-scale customizations for environmental observatories. Ongoing efforts have been focusing on further developing ubiquitous provenance service from sensor to streaming data processing, cyberenvironment mashup capabilities, and dynamic evolution of terminology and usage of multiple ontologies.

7. ACKNOWLEDGMENTS

Our thanks go to the entire NCSA ECID project team. Funding for ECID technology development comes from the National Science Foundation (BES-0414259, BES-0533513, and SCI-0525308) and the Office of Naval Research (N00014-04-1-0437).

8. REFERENCES

[1] J. Myers, T. Dunning, Cyberenvironments and Cyberinfrastructure: Powering Cyber-research in the 21st Century, Proceedings of the Foundations of Molecular Modeling and Simulation, Blaine, WA, July 9-14, 2006. Available: http://cet.ncsa.uiuc.edu/CEResources/background/FOMMS_MyersDunning.pdf

[2] Water Science and Technology Board (WSTB), Division on Earth and Life Studies, National Research Council of the National Academies. (2006). CLEANER and NSF's Environmental Observatories. National Academy Press. 2006. Available: http://orsted.nap.edu/openbook.php?record_id=11657&page=R1

[3] B. Minsker, E. Coopersmith, B. Hodges, D. Maidment, J. Bonner, and P. Montagna. An Environmental Information System for Hypoxia in Corpus Christi Bay: A WATERS Network Testbed. Presented at AGU 2006 Fall Meeting, San Francisco, CA, December 11-15, 2006.

[4] myGrid. (2006). myExperiment. [Online]. http://myexperiment.org/

[5] B, Minsker,, J. Myers, M. Marikos, T. Wentling, S. Downey, Y, Liu, P. Bajcsy, R. Kooper, L. Marini, N. Contractor, H. Green, Y. Yao, J. Futrelle. Environmental CyberInfrastructure Demonstrator Project: Creating Cyberenvironments for Environmental Engineering and Hydrological Science Communities. Presented at Supercomputing Conference 2006 (SC06), Tampa, FL, November 13-17, 2006.

[6] Y. Liu, S. Downey, B. Minsker, J. Myers, T. Wentling, and L. Marini, "Event-Driven Collaboration through Publish/Subscribe Messaging Services for Near-Real- Time Environmental Sensor Anomaly Detection and Management," Eos Trans. AGU 87(52), Fall Meet. Suppl. 2006.

[7] L. Marini, B. Minsker, R. Kooper, J. Myers, and P. Bajcsy, "CyberIntegrator: A Highly Interactive Problem Solving Environment to Support Environmental Observatories", Eos Trans. AGU 87(52), Fall Meet. Suppl. 2006.

[8] H. D. Green, N. S. Contractor, and Y. Yao, "CI-KNOW: Cyberinfrastructure Knowledge Networks on the Web. A Social Network Enabled Recommender System for Locating Resources in Cyberinfrastructures", Eos Trans. AGU 87(52), Fall Meet. Suppl. 2006.

[9] J. Futrelle, J. Myers, B. Minsker, and P. Bajcsy. Community-based Metadata Integration for Environmental Research. Presented at the Seventh International Conference on Hydroscience and Engineering (ICHE-2006), Philadelphia, PA, September 10-13, 2006.

[10] Consortium of Universities for the Advancement of Hydrologic Science. (2006). Hydrologic Observations Data Model. [Online]. http://www.cuahsi.org/his/tk-observedb.html