Environmental Modelling & Software 51 (2014) 59–69

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Research paper

A centralized tool for managing, archiving, and serving point-in-time data in ecological research laboratories

Seth J.K. Mason [a,*], Sean B. Cleveland [b], Pol Llovet [b], Clemente Izurieta [a,d], Geoffrey C. Poole [c,b,d]

[a] Department of Computer Science, Montana State University, 357 EPS Building, Bozeman, MT 59717, USA
[b] Research Computing Group, Montana State University, P.O. Box 173505, Bozeman, MT 59715, USA
[c] Department of Land Resources and Environmental Sciences, Montana State University, P.O. Box 173120, Bozeman, MT 59717, USA
[d] Montana Institute on Ecosystems, Montana State University, 106 AJM Johnson Hall, Bozeman, MT 59717, USA

ARTICLE INFO

Article history:
Received 13 September 2012
Received in revised form
30 August 2013
Accepted 11 September 2013
Available online 18 October 2013

Keywords:
Data management
Cyberinfrastructure
Data models
Data schemas
Informatics

ABSTRACT

The recent proliferation of software tools that aid researchers in various phases of data tracking and analysis undoubtedly contribute to successful development of increasingly complex and data-intensive scientific investigations. However, the lack of fully integrated solutions to data acquisition and storage, quality assurance/control, visualization, and provenance tracking of heterogeneous temporal data streams collected at numerous geospatial locations continues to occupy a general problem area for scientists and data managers working in the environmental sciences. We present a new Service Oriented Architecture (SOA) that allows users to: 1) automate the process of pushing real-time data streams from networks of environmental sensors or other data sources to an electronic data archive; 2) to perform basic data management and quality control tasks; and 3) to publish any subset of the data to existing cyberinfrastructure platforms for global discovery and distribution via the World Wide Web. The approach outlined here supports management of: 1) repeated field observations, 2) data from laboratory analysis of field samples, 3) simulation results, and 4) derived values. We describe how the use of Hypertext Transfer Protocol (HTTP) Application Programming Interfaces (APIs) Representational State Transfer (REST) methods for data model objects and Resource Query Language (RQL) interfaces respond to a basic problem area in environmental modelling by enabling researchers to integrate an electronic data repository with existing workflows, simulation models, or third-party software.

1. Introduction

The fields of science and engineering are expected to witness the collection of unprecedented quantities of data in the coming decades as sensor technologies become more affordable and remote or autonomous data acquisition networks become more prolific. Developments in data collection drive increasingly data-intensive research efforts in the environmental sciences. Advances in data generation and computational performance herald changes in the way that some scientific research is conducted and utilized, yet realization of their full potential depend on concurrent efforts to develop supporting cyberinfrastructure.

Cyberinfrastructure integrates computing hardware, digitally enabled sensors, data observatories and experimental facilities, interoperable software and middleware services and tools, and data and networks (National Science Foundation, 2007). The needs for cyberinfrastructure are diverse and often require domain-specific solutions. Notable recent cyberinfrastructures developed to support research activities of individuals and research groups in the environmental sciences, include DataONE (www.dataone.org), the Long Term Ecological Research Network (LTER) (Karasti et al., 2006), and the Consortium of Universities for the Advancement of Hydrologic Sciences Inc., Hydrologic Information System (CUAHSI-HIS) (Horsburgh et al., 2009). The data publication systems at the core of each of these efforts significantly enhance researchers' ability to conduct research with data covering large spatial extents, ease the discovery of diverse datasets, and facilitate collaborative research.

Advances in data generation, publication, and discovery make possible many new types of investigation; however, the outcome of

* Corresponding author. Tel.: +1 970 903 7561; fax: +1 406 994 7438.
E-mail addresses: seth.kurtmason@msu.montana.edu, seth.k.mason@gmail.com (S.J.K. Mason).

1364-8152/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.envsoft.2013.09.008

any given research effort and the value of any published data set hinges largely on an investigator's effectiveness at maintaining a high level of data control while steering data generated by electronic sensors in the field or laboratory through quality assurance and quality control (QA/QC) procedures, visualization, data analysis, integration with simulation models and, ultimately, publication. The task of shepherding data through the various components of the data life cycle is herein referred to as '*data management*'. As the quantity and diversity of data both generated by research groups and required by increasingly complex analyses grows, the task of identifying effective strategies for comprehensive data management becomes progressively more arduous (Pokorny, 2006). The burden of identifying end-to-end data management solutions that both satisfy domain-specific and lab-specific requirements typically falls to the individual investigator. As a result, researchers may resort to utilization of multiple data management support systems or applications (each requiring some non-trivial amount of training for proper use), or simply relying on raw data files to support all archival and analysis needs. In recognition of the significant shortcomings of the latter approach, the National Science Foundation (NSF) now requires funded researchers to detail and implement data management plans.

If the ultimate goals of cyberinfrastructure development include increasing the availability of diverse data sets, facilitating collaborative and interdisciplinary research, and, thus, enabling scientific discoveries (Horsburgh et al., 2009), then the integration of data management support software tools and data publication and discovery systems warrants further attention. Solutions developed for other domain sciences provide a valuable template for such development efforts. The iPlant cyberinfrastructure developed to support bioinformatics in the plant sciences provides data discovery, analysis, mining, and visualization tools through a common web interface. In the environmental sciences, both DataONE and CUAHSI-HIS provide investigators with a number of independent desktop applications (e.g. HydroDesktop for CUAHSI-HIS and ONEDrive for DataONE), plugins for commonly used software packages (e.g. HydroGET for CUAHSI-HIS and ONE*R* for DataONE), and web applications (e.g. HydroServer Map for CUAHSI-HIS and ONEMercury for DataONE) that support the core publication system. End-to-end data management solutions developed in support of LTER require the use of software tools like the GCE Data Toolbox for Matlab (https://gce-svn.marsci.uga.edu/trac/GCE_Toolbox) or customized development of a web-based metadata management support system using packages like DEIMS (http://code.google.com/p/deims/). Such tools provide investigators with the means for conducting QA/QC procedures, visualizing and analysing data, or incorporating data into scientific workflows or simulation models. We suggest that while, in aggregate, these systems and the array of supporting software and web tools currently available to researchers in the environmental sciences represent a powerful set of tools. However, the relative lack of availability of open-source, *integrated* data acquisition, collation, data management support and publication solutions limits both the quantity and quality of data that would otherwise be destined for an electronic publication node and, thereby, made readily available to further advance human knowledge. Failure or inability of cyberinfrastructure to support the breadth of data management activities constrains researchers' ability and willingness to contribute high-quality data to existing electronic publication and discovery systems.

Here we present a description of the Data Hub for the Virtual Observatory and Ecological Informatics System (the "VOEIS Data Hub") as a specific instance of a solution to a general problem in the environmental sciences: the use of a centralized data storage system by loosely affiliated or unaffiliated data producers for the compilation, management, and redistribution of data sets. Key aspects of the system include the ability for individual researchers to amass and manage large, continually growing, controlled access data sets, while also providing a "point-and-click" solution for sharing any subset of their data publically via the World Wide Web. The VOEIS system, or aspects of its implementation may be useful to businesses, governments, NGOs, or other organizations that collect and manage geographically distributed information, a portion of which is destined for publication at some point after data collection.

1.1. Motivations for development

1.1.1. Hosting and administration

Existing cyberinfrastructure significantly leverages the power of the Internet to enable data sharing and discovery. Systems like CUAHSI-HIS and DataONE provide users with web services for accessing data and with software components for performing some aspects of data management over the web. Historically, these systems relied on a federated approach to server administration where individual users or research groups set up and administered their own servers (i.e. HydroServers and Member Nodes) for data archival and publication. Conversely, an approach to cyberinfrastructure development that utilizes the Software as a Service (SaaS) delivery model partially alleviates the requirements for capital and access to IT expertise and support associated with software development or server maintenance. Cyberinfrastructure utilizing the SaaS model can be supported by a group of researchers, by a department, or by an academic or governmental institution. Realization of savings from this approach stems from defrayed costs—administration of a single data archival, publication and data management support system (hosted on a single server) serving many independent research efforts. Critically, cyberinfrastructure solutions utilizing the SaaS delivery model must provide researchers with enhanced user access controls to data repositories, as data from multiple research groups resides on a common server.

1.1.2. User access

One of the chief motivating factors in the development of VOEIS was the conspicuous absence of access restriction features in many existing cyberinfrastructures. Open data policies raise privacy issues for those users confronted with legal restrictions on use and application of their data, dealing with distrust among collaborators, or concerned with issues of professional credit or rights to data use (Steiner et al., 2009). As a result, researchers who wish to separate public and private data must utilize two separate data archival and publication solutions: a private, local system and a public one. This situation creates some obvious difficulties for those researchers with resource or technical constraints that prevent either the administration of multiple data archival/management systems or the development of custom software tools. At time of writing, the authors were unaware of the existence of any ready-made and publically available solutions that address this problem. Thus, addressing these concerns requires development of a data management solution that exhibits highly granular user access controls while simultaneously supporting publication to existing data publication nodes (e.g. a CUAHSI-HIS HydroServer).

1.1.3. Quality assurance and quality control

The capacity of scientific research findings to contribute to human knowledge and advance basic science depends largely on the quality of data used to generate those findings. Data quality can be degraded by errors introduced at various points in the data life cycle: during data generation due to sensor malfunction, during data formatting and archival due to software problems, and during processing and analysis as a result of user error. As rates of data