Environmental Modelling & Software 54 (2014) 73–87

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

Sustainable long term scientific data publication: Lessons learned from a prototype Observatory Information System for the Illinois River Basin

Benjamin L. Ruddell [a,*], Ilya Zaslavsky [b], David Valentine [b], Bora Beran [c], Michael Piasecki [d], Qingwei Fu [e], Praveen Kumar [f]

[a] Arizona State University, Tempe, AZ, USA
[b] San Diego Supercomputing Center, University of California San Diego, La Jolla, CA, USA
[c] Microsoft Research, Redmond, WA, USA
[d] The City College of New York, Department of Civil Engineering, New York, NY, USA
[e] Brookfield Renewable Energy Group, Marlborough, MA, USA
[f] University of Illinois at Urbana-Champaign, Department of Civil and Environmental Engineering, Urbana, IL, USA

ARTICLE INFO

Article history:
Received 26 April 2013
Received in revised form
19 November 2013
Accepted 23 December 2013
Available online 19 January 2014

Keywords:
Cyberinfrastructure
Observatory Information System
Metadata
Curation
Redistribution
Federated
Hydrology
Hydroinformatics
River basin
Geoscience
Long-term
Sustainable
Standards
ISO-19115
FGDC
XML

ABSTRACT

In 2005 a prototype Observatory Information System (OIS) was developed for the Illinois River Basin Observatory (IRBO), connected to a federated scientific data network, populated with a representative collection of legacy datasets, and linked to external data streams. The perspective of seven years' time and the disestablishment of the system provide an opportunity to study the system life cycle. We detail best practices for multi-level OIS design for long-term performance, based on a publication-mandatory metadata implementation standard using ISO-19115. These principles balance general users' requirements against the requirements of specific scientific applications, and maximize the system's capacity to deal with legacy and heterogeneous data sources, enhancing long-term sustainability and flexibility for diverse multi-level user groups. These findings are relevant to ongoing developments of networked Scientific Information Systems that are increasingly critical to support and sustain the long-term benefits of modeling and observatory science.

1. Introduction

Borgman et al. (1996) describe the Information Life Cycle of data as, Creation, Contextualization, Search, and Utilization. The ideal system for data and model publication and re-use would be a component of a research project from the start, in reality it is often introduced just before the end of the typical project's Information Life Cycle to prevent data from being lost at the end of the project. Data publication systems facilitate "long-lived" or "permanent" access to scientific data (Hodge and Frangakis, 2005). In the past, little data sharing has occurred in science, owing to "…lack of demand, lack of standards, and concerns about publication, ownership, data quality, and ethics." (Borgman et al., 2006). This is a general problem for 21st century science, and it begs for a general solution.

* Corresponding author.
E-mail addresses: bruddell@asu.edu (B.L. Ruddell), zaslavsk@sdsc.edu (I. Zaslavsky), valentin@sdsc.edu (D. Valentine), borabe@microsoft.com (B. Beran), mpiasecki@ccny.cuny.edu (M. Piasecki), Qingwei.fu@brookfieldrenewable.com (Q. Fu), kumar1@illinois.edu (P. Kumar).

1364-8152/$ – see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.envsoft.2013.12.015

This application area blends library and information science with software engineering, in the context of data-driven science. The area is barely a generation old. No mature consensus on the specific design principles of a universal scientific data and model publication system has yet emerged. Instead, each scientific subdiscipline and community of practice has made efforts to develop its own software systems (e.g. Borgman et al., 2006, or Nambiar et al., 2006). Some of these systems have survived to become de-facto disciplinary community standards. The ecological and geoscience communities have become leaders in rigorously systematizing these applications, perhaps owing to the unique complexity of data and models in these empirical sciences (e.g. Peckham et al., 2013; Mason et al., 2014; Nativi et al., 2013; Brooking and Hunter, 2013; Goodall et al., 2008; Maidment et al., 2009; Horsburgh et al., 2010; Zaslavsky et al., 2012; Karastl et al., 2006; KNB, 2013; Helly et al., 1999; Rizzoli et al., 1998; FGDC, 1998).

Over time, the successful disciplinary initiatives have begun to cooperatively seek broader and more unified solutions to the problem of data and model publication. For example, this is a goal of the U.S. National Science Foundation's DataNet program which aims to, "provide reliable digital preservation, access, integration, and analysis capabilities for science and/or engineering data over a decades-long timeline" (NSF, 2007a,b). Several current programs, including DataONE (Michener et al., 2011a,b) and EarthCube (NSF, 2013), share this aim. Experience is making it clear that best practices exist to address the problems and challenges of a generalized data and model publication system, to maximize sustainability and utility for the broadest possible community of users.

The term 'cyberinfrastructure' (CI), "refers to infrastructure based upon distributed computer, information and communication technology" (Atkins et al., 2003). An issue of current interest is the development of scientific CI to facilitate multidisciplinary, multi-user science in the context of environmental observatories and the models that utilize this data. Two examples of such observatories include the National Ecological Observation Network (NEON) and the Critical Zone Observatories (CZO). This CI is necessary to allow scientists to organize and publish their data in a manner accessible to other scientists within and outside the immediate spatial and topical research domain (Peters et al., 2008). The CI bridges the gap between the creators and potential users of data and models, and enables data preservation and re-use. This is made possible by balancing the needs of both specific domain user communities and by the general user community. The publishing user needs low overhead for data publication, while maintaining application-specific metadata information needed for immediate science applications. Meanwhile, a user from the general scientific community needs to discover, download, and utilize resources from a diverse set of publishers with minimal time and effort (Michener, 2006). The CI thereby enables synthetic and integrated interdisciplinary scientific study and modeling in the "natural laboratory" facilitated by scientific observatories that collect environmental data (CIF21, 2012). The creation of observatory information systems and digital libraries is now helping to spur a sea-change in scientific data sharing, by providing the empirical basis for the scientific discoveries that so often occur at the boundaries between disciplines (Lemaine et al., 1976). Clear definitions of the purpose of Scientific Information Systems (SIS) may not yet have reached a consensus, but patterns are emerging. Helly et al. (1999) draws a distinction between "digital libraries", which emphasize metadata for preservation, archival and distribution of data in a fashion that is application-agnostic, and "data repositories", which emphasize the curation of data content related to a particular scientific application. Environmental Observatory Information Systems (EOIS/OIS) are a specific type of SIS that has been generally defined as facilitating functions including, "collection, organization, storage, analysis, and publication of environmental observations" (Horsburgh et al., 2011). An OIS therefore incorporates both digital library and data repository purposes, supporting both general data publication and specific scientific applications using the data.

Digital data publication is more than simply 'cataloging' data and metadata to make a dataset searchable and available to users; it sometimes implies a thorough process of 'curation' which involves reprocessing, documentation, quality control, metadata (re)construction, and the production of derivative products. Cataloging and curation permit users to utilize and properly cite the data products of the observatory (Michener, 2006; Helly, 2006). Federated systems allow for more general requirements to be met at a higher level by networking a number of lower level systems (Sheth and Larson, 1990). The design of the OIS's metadata implementation standard is therefore particularly crucial for a multi-level OIS's functionality because it both enables and limits specific types of cataloging and curation problems that can be solved at higher levels of the federated OIS. A good metadata standard provides the flexibility and generality required for data synthesis in contexts outside of the specific scientific applications for which the data was originally collected (Beran and Piasecki, 2009; Colomb, 1997; Gray et al., 2005).

Current high-level scientific information system efforts leverage prior development efforts to create an integrated vision for a cross-domain CI for the geosciences. A recent whitepaper produced to inform the U.S. National Science Foundation's EarthCube scientific CI funding program emphasized broad community engagement, system-level governance, reliance on metadata and data standards and formal and explicit information models, support of provenance, data quality description, and reliable long-term data access and re-use as critical components of establishing trust in data contributed by other stakeholders (Zaslavsky et al., 2012). This paper addresses federated OIS models and metadata systems that are compatible with these broad priorities for science CI.

Our oldest environmental observatories are merely decades old, but are gaining initial experience solving long-term digital information management problems for environmental observatories (e.g. Karasti et al., 2006; Michener et al., 2011a,b). The problem of 'digital obsolescence' of data hosted and published by OIS's is becoming more severe as new OISs are built, while at the same time many small-scale or project-specific OIS project life cycles will be ending (National Science Board, 2005). A recent example is the National Biological Information Infrastructure (NBII, 2012), a major digital library that was defunded in January of 2012. How will legacy datasets be preserved and continue to be re-usable in the future, and are current information system strategies plausibly successful toward this end? By following the implementation of an OIS prototype that was designed to incorporate legacy data into the broader federated digital library network, and which then itself was taken offline, it is possible to gain valuable retrospective information on the possible fate of our OIS's. This information will help us both to design OIS's to integrate legacy data and to prepare for the time when today's OIS's will be disestablished or replaced with fundamentally different or more advanced technologies.

In 2004–2005, work was completed by the authors on the prototype IRBO-OIS. This paper provides a unique retrospective insight into the best practices and pitfalls of developing an OIS and the underlying OIS metadata standards, including criteria of generality and sustainability of the CI methods chosen, from the vantage point seven years after the work was conducted. The