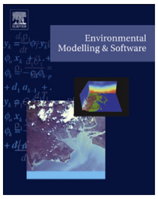
Environmental Modelling & Software 79 (2016) 55–74

Contents lists available at ScienceDirect

Environmental Modelling & Software

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



Observations Data Model 2: A community information model for spatially discrete Earth observations

Jeffery S. Horsburgh [a, *], Anthony K. Aufdenkampe [b], Emilio Mayorga [c], Kerstin A. Lehnert [d], Leslie Hsu [d], Lulin Song [d], Amber Spackman Jones [e], Sara G. Damiano [b], David G. Tarboton [e], David Valentine [f], Ilya Zaslavsky [f], Tom Whitenack [f]

[a] *Department of Civil and Environmental Engineering, Utah Water Research Laboratory, Utah State University, 8200 Old Main Hill, Logan, UT 84322-8200, USA*
[b] *Stroud Water Research Center, 970 Spencer Road, Avondale, PA 19311, USA*
[c] *Applied Physics Laboratory, University of Washington, 1013 NE 40th Street, Seattle, WA 98105-6698, USA*
[d] *Lamont-Doherty Earth Observatory, Columbia University, 61 Route 9W, Palisades, NY 10964-8000, USA*
[e] *Utah Water Research Laboratory, Utah State University, 8200 Old Main Hill, Logan, UT 84322-8200, USA*
[f] *San Diego Supercomputer Center, University of California San Diego, 9500 Gilman Drive #0505, La Jolla, CA 92093-0505, USA*

ARTICLE INFO

Article history:
Received 1 July 2015
Received in revised form
29 January 2016
Accepted 29 January 2016
Available online xxx

Keywords:
Observations
Information model
Data management
Interoperability
Cyberinfrastructure
Data publication

ABSTRACT

Integrated access to and analysis of data for cross-domain synthesis studies are hindered because common characteristics of observational data, including time, location, provenance, methods, and units are described differently within different information models, including physical implementations and exchange schema. We describe a new information model for spatially discrete Earth observations called the Observations Data Model Version 2 (ODM2) aimed at facilitating greater interoperability across scientific disciplines and domain cyberinfrastructures. ODM2 integrates concepts from ODM1 and other existing cyberinfrastructures to expand capacity to consistently describe, store, manage, and encode observational datasets for archival and transfer over the Internet. Compared to other systems, it accommodates a wider range of observational data derived from both sensors and specimens. We describe the identification of community information requirements for ODM2 and then present the core information model and demonstrate how it can be formally extended to accommodate a range of information requirements and use cases.

Software availability

Name of software: Observations Data Model 2 (ODM2)
Developers: Jeffery S. Horsburgh, Anthony K. Aufdenkampe, Emilio Mayorga, Kerstin A. Lehnert, Leslie Hsu, Lulin Song, Amber Spackman Jones, Sara G. Damiano, David G. Tarboton, David Valentine, Ilya Zaslavsky, Tom Whitenack
Contact: Jeffery S. Horsburgh; Address: 8200 Old Main Hill, Logan, UT 84322-8200, USA; Email: jeff.horsburgh@usu.edu
Year first available: 2015
Required hardware and software: ODM2 is available for use with Microsoft SQL Server, MySQL, PostgreSQL, and SQLite on Windows, Macintosh, and Linux based computers. Information about additional software available for working with ODM2 is available at https://github.com/ODM2/ODM2.
Cost: Free. Software and source code are released under the New Berkeley Software Distribution (BSD) License, which allows for liberal reuse. All source code, examples, and documentation can be accessed at https://github.com/ODM2/ODM2.

---

* Corresponding author.
E-mail addresses: jeff.horsburgh@usu.edu (J.S. Horsburgh), aufdenkampe@stroudcenter.org (A.K. Aufdenkampe), mayorga@apl.washington.edu (E. Mayorga), lehnert@ldeo.columbia.edu (K.A. Lehnert), lhsu@ldeo.columbia.edu (L. Hsu), song@ldeo.columbia.edu (L. Song), amber.jones@usu.edu (A.S. Jones), sdamiano@stroudcenter.org (S.G. Damiano), dtarb@usu.edu (D.G. Tarboton), valentin@sdsc.edu (D. Valentine), zaslavsk@sdsc.edu (I. Zaslavsky), twhitenack@sdsc.edu (T. Whitenack).

http://dx.doi.org/10.1016/j.envsoft.2016.01.010
1364-8152/

1. Introduction

As volumes of Earth observations increase, so does the importance of their efficient management and use. In the past several years, a number of cyberinfrastructures have emerged for sharing spatially discrete Earth observations data, including the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) Hydrologic Information System (HIS) (Tarboton et al., 2009), the Critical Zone Observatory Integrated Data Management System (CZOData) (Zaslavsky et al., 2011), the Integrated Earth Data Applications (IEDA) and EarthChem system (Lehnert et al., 2011a, 2004, 2009), and the Integrated Ocean Observing System (IOOS) (IOOS, 2010a; Lubchenco, 2010). These systems are built using the principles of service-oriented architecture (SOA) (Josuttis, 2007; Goodall et al., 2008) and rely on standard data encodings and, in some cases, standard semantics for classes of geoscience data. A core focus of most of these systems is on publishing or sharing data on the Internet via web services and domain specific encodings or markup languages.

While these systems have made considerable progress in making data interoperable and available, it still takes a knowledgeable investigator substantial effort to discover and access datasets from multiple domain-specific repositories for analysis because of inconsistencies in the way the different domain systems describe, encode, and share data. First, data structures used by existing domain cyberinfrastructures are often insufficient to store or describe the entire range of Earth observations. Here we refer to the sufficiency of metadata with respect to both data discovery and ultimate use. For example, data structures and encodings used by the CUAHSI HIS contain the necessary metadata to describe time series of *in situ* observations made at point locations such as streamflow gages and weather stations. However, they are inadequate for water quality or solid Earth geochemical samples taken in the field and analyzed later in a laboratory because existing method and sample descriptions do not contain all of the needed metadata and are not extensible to allow, for example, important data structures such as sample fractions and sub-sample parent–child relationships. Conversely, the EarthChem system contains the necessary metadata elements and structures to effectively describe observations derived from *ex situ* analysis of geochemical samples, but is not well structured to support time series of observations from *in situ* sensors.

Yet, there are many research scenarios that require efficient integration of these data types across different domains of observational Earth science. For example, understanding a soil profile's geochemical response to extreme weather events requires integration of hydrologic and atmospheric time series with geochemical data from soil sample fractions collected over various depth intervals from soil cores or pits at different positions on a landscape. Similarly, understanding spatial and temporal patterns in suspended sediment fluxes, sources, and associated contaminants in response to land use and climate change requires close integration of hydrologic time series with a variety of geochemical data analyzed in different laboratories on separate sample fractions (e.g., acid extract of fine sediments for heavy metals, solvent extract of whole water for organic contaminants, dried filter for suspended solids concentration). Currently, integrated access to and analysis of data for such studies are hindered because common characteristics of observational data, including time, location, provenance, methods, and units are described using different constructs within different systems. Integration requires multiple syntactic and semantic translations that are, in many cases, manual, error-prone, and/or lossy. Management of data across multiple repositories/systems is similarly complicated, and data managers such as those managing diverse datasets from large projects like Critical Zone Observatories may prefer a single schema that enables a more efficient data integration strategy (e.g., more straightforward, lossless, sustainable, reusable, etc.) rather than managing multiple different domain databases with different schemas.

While there are many properties of observations that are common across the various types of observational data acquired and used within the geosciences, each domain also presents observation types that are unique. In many instances, data structures have been built to support the most common types of observations within a specific domain, without consideration of the broader context of available observations across domains, leading to substantial syntactic and semantic heterogeneity in observational data representations. Semantic and syntactic heterogeneity are major hurdles to be overcome, especially across data types and scientific domains (Beran and Piasecki, 2009; Horsburgh et al., 2009; Hankin et al., 2010a, 2010b). Because many systems (including those mentioned above) already have their own existing data structures and/or database implementations, one solution to achieving interoperability between systems is to agree upon a common information model to which data in the existing systems can be mapped. The common element, then, across systems and the services that they provide is the information model, with physical implementations within various file systems and databases for data storage, within extensible markup language (XML) schemas and file formats for data transfer, and within web service interfaces that provide access to data. Indeed, overcoming heterogeneity and achieving interoperability and more reliable data integration within SOAs depends on standardizing descriptions of common characteristics within a common information model and well-defined interfaces and data encodings that implement it.

An information model is a representation of concepts, relationships, constraints, rules, and operations that specify the semantics of data for a chosen domain of discourse (Lee, 1999). At its simplest level, an information model defines the domain's entity types and their properties, relationships, and allowed operations on the entities. In relational database design terminology, an information model is essentially equivalent to a “conceptual database model” (e.g., Connolly and Begg, 2005). In a relational database implementation of an information model, entities become tables and their properties become table columns. More generally, an information model provides a sharable, stable, and organized structure of the information requirements for a domain context, without constraining how that description is mapped to an actual implementation in software (Fulton, 2006). There may be many mappings of the information model. Such mappings are called data models, irrespective of whether they are object models, entity relationship models such as those used by relational databases, or XML schemas. The fundamental elements within the information model are based on the domain of discourse to be described and how the model will be used – e.g., to support data discovery or data storage. For example, a rich set of descriptive metadata about the variables that were observed and the context within which an observation was made is fundamental for both discovering and interpreting observational data (Madin et al., 2007). The information model behind a data system is thus critically important to the effectiveness and interoperability of any cyberinfrastructure.

Domain-agnostic information models for observations have been developed and standardized (e.g., the Open Geospatial Consortium's (OGC) Observations and Measurements (O&M) standard (Cox, 2010)). While they provide a general framework and key constructs for describing different types of observations, these models are expected to be used as a common basis for domain-specific profiles of information exchange. In this paper, we demonstrate how a single, detailed profile can be developed for a wider range of geoscience domains. As architects and developers of