DATA SCIENCE JOURNAL

Hsu, L et al 2017 Enhancing Interoperability and Capabilities of Earth Science Data using the Observations Data Model 2 (ODM2). *Data Science Journal,* 16: 4, pp. 1–16, DOI: https://doi.org/10.5334/dsj-2017-004

RESEARCH PAPER

Enhancing Interoperability and Capabilities of Earth Science Data using the Observations Data Model 2 (ODM2)

Leslie Hsu[1], Emilio Mayorga[2], Jeffery S. Horsburgh[3], Megan R. Carter[1], Kerstin A. Lehnert[1] and Susan L. Brantley[4]

[1] Lamont-Doherty Earth Observatory, Columbia University, 61 Route 9W, Palisades, NY 10964, USA

[2] Applied Physics Laboratory, University of Washington, 1013 NE 40th Street, Seattle, WA 98105-6698, USA

[3] Department of Civil and Environmental Engineering and Utah Water Research Laboratory, Utah State University; 8200 Old Main Hill, Logan, UT 84322-8200, USA

[4] Department of Geosciences, Pennsylvania State University, University Park, PA 16802, USA

Corresponding author: Leslie Hsu (lhsu@usgs.gov)

Earth Science researchers require access to integrated, cross-disciplinary data in order to answer critical research questions. Partially due to these science drivers, it is common for disciplinary data systems to expand from their original scope in order to accommodate collaborative research. The result is multiple disparate databases with overlapping but incompatible data. In order to enable more complete data integration and analysis, the Observations Data Model Version 2 (ODM2) was developed to be a general information model, with one of its major goals to integrate data collected by *in situ* sensors with those by *ex-situ* analyses of field specimens. Four use cases with different science drivers and disciplines have adopted ODM2 because of benefits to their users. The disciplines behind the four cases are diverse – hydrology, rock geochemistry, soil geochemistry, and biogeochemistry. For each case, we outline the benefits, challenges, and rationale for adopting ODM2. In each case, the decision to implement ODM2 was made to increase interoperability and expand data and metadata capabilities. One of the common benefits was the ability to use the flexible handling and comprehensive description of specimens and data collection sites in ODM2's sampling feature concept. We also summarize best practices for implementing ODM2 based on the experience of these initial adopters. The descriptions here should help other potential adopters of ODM2 implement their own instances or to modify ODM2 to suit their needs.

Keywords: observations; information model; data management; interoperability; cyberinfrastructure

Introduction

The magnitude and diversity of Earth science observations is increasing exponentially. This data richness is fueling new and novel studies that advance our understanding of Earth and environmental processes; however, tools for integrating, efficiently managing, properly documenting, and making such data accessible to diverse groups of collaborating scientists are still in early stages. The need for such tools is clearly shown by large collaborative projects such as the Critical Zone Observatories (e.g., Brantley et al., 2007). These projects include many investigators that produce many different types of data, much of which is often stored and managed in its own domain-specific data system or by idiosyncratic, custom approaches. While domain-specific data systems may provide advanced functionality for particular data types, working with them requires disciplinary knowledge for navigation and data access. The specificity of data types within these systems has commonly led to challenges in integrating data across systems, domains, and research groups. To achieve the science goals of these collaborative projects, however, the different data types must be compared, integrated, and analyzed across sources.

This need for interdisciplinary data interoperability across scientific disciplines and domain cyberinfrastructures drove the development of the Observations Data Model Version 2 (ODM2) (Horsburgh et al., 2016). ODM2 integrates concepts from ODM1 (Horsburgh et al., 2008) and other existing geoscience cyberinfrastructures to expand capacity to consistently describe, store, manage, and encode observational datasets for archival and transfer over the Internet. ODM2's core schema, inspired by and developed as a profile of the Open Geospatial Consortium's Observations & Measurements (O&M) standard's general concept of observations (Cox, 2007a; 2007b; 2011), provides a consistent way to describe Earth observations of many different types. O&M was the original source of many of the fundamental concepts in the ODM2 data model that motivated different users to adopt ODM2.

Here we describe several data use cases that illustrate the challenges that accompany geoscience datasets and the benefits of using an advanced information model like ODM2. The disciplines behind the cases are diverse – hydrology, rock geochemistry, soil geochemistry, and river biogeochemistry. We describe how ODM2 enables consistent description of common elements of diverse data types while still enabling more specific data structures for each type, thus accommodating the varied requirements for each use case. We also describe how use case requirements led to design choices for ODM2 that enhanced buy-in from multiple adopters. These examples show how a general model can accommodate diverse requirements. We anticipate that this work will help other potential adopters of ODM2 implement their own instances or modify ODM2 to suit their needs.

Background

In this section we provide a brief overview of the ODM2 information model as background for the use case implementations that follow. For a more comprehensive description, readers are directed to Horsburgh et al. (2016). Although the ODM2 information model (**Figure 1**) is not specific to any particular physical implementation (e.g., relational database, XML schema, etc.), we have chosen to illustrate it here using standardized entity relationship notation to describe its core functionality and because each of the use cases that follow were implemented and tested using relational databases.

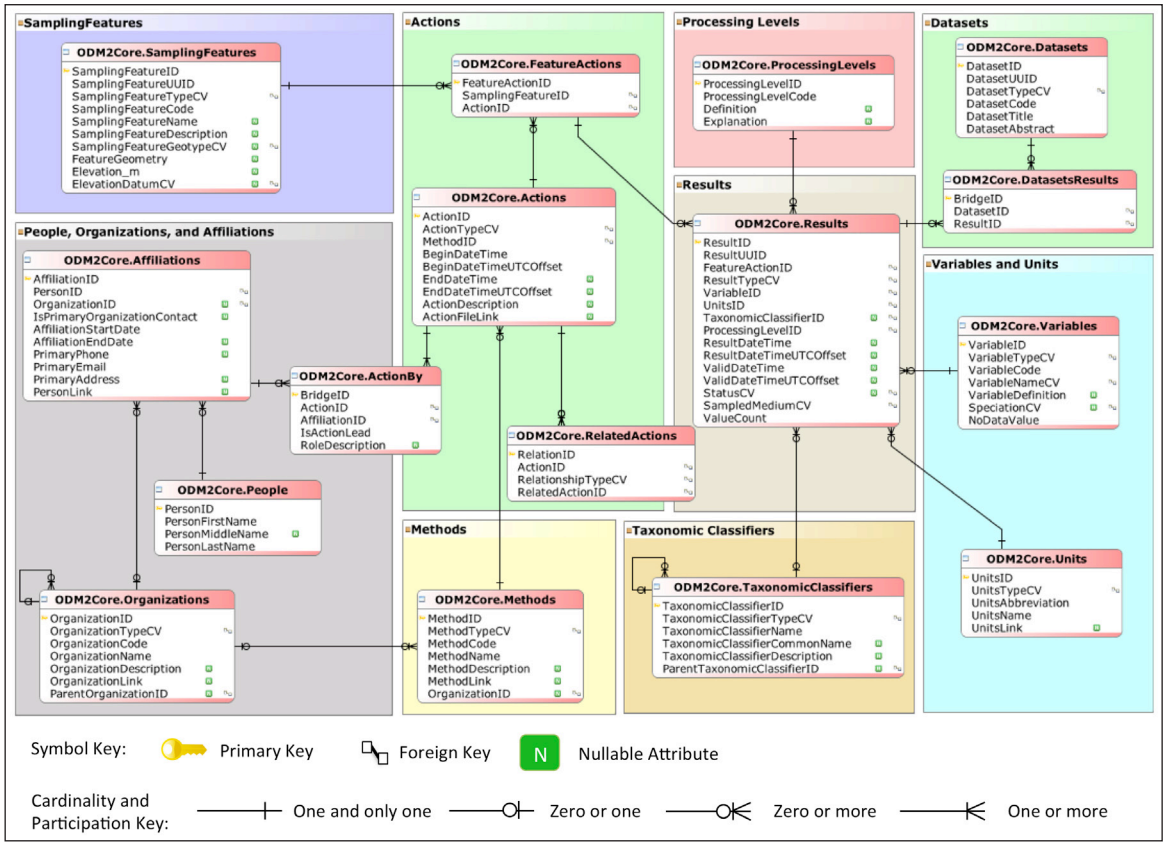

Figure 1: ODM2 core schema illustrated using entity relationship notation (Horsburgh et al., 2016).

ODM2 is organized with a core schema (**Figure 1**) and multiple extension schemas (not shown here, but e.g. Annotations, Data Quality, Equipment, External Identifiers, and Results. Full list at https://github.com/ODM2/ODM2/wiki/Documentation). The core schema contains entities, attributes, and relationships that are common for all observations, regardless of their type. Extension schemas add functionality to ODM2 by adding entities, attributes, and relationships needed for particular use cases or data types. Every implementation of ODM2 uses the core schema, but use of extension schemas may vary by use case, as described in the cases below. In ODM2, an "observation" is made up of two elements: an Action performed on or at a SamplingFeature that produces an observation Result, and a Result that is the outcome of that Action. These important concepts were adopted from both OGC's O&M standard (Cox, 2011) and from laboratory data models (e.g., Wendl et al., 2007). The separation of Actions and their Results enables a single Action to produce multiple Results, Actions of many types that may or may not produce results (e.g., a maintenance Action versus an instrument deployment Action), and Results that may be of multiple different types (e.g., individual laboratory measurements versus time series of sensor observations). Horsburgh et al. (2016) provide detailed descriptions of how Action descriptions can be combined to form sampling or sensor workflows. Variables, Methods, ProcessingLevels, and Units are largely based on ODM1, but their ODM2 implementation enables more refined distinctions. A more sophisticated model of People and Organizations was implemented in ODM2 to facilitate better descriptions of who performed specific Actions.

In the sections that follow, we illustrate how the specific requirements of multiple geoscience data use cases were mapped to the ODM2 core schema and its extensions. These sections describe the benefits of using ODM2 in terms of organization, support for capturing required metadata, and advances made over how the data were previously stored and/or managed. We also describe challenges that were faced in adopting ODM2 related to mapping of metadata between different models and other semantic challenges, complexity and granularity of metadata, and any technical challenges that had to be overcome.

Use cases

The use cases described here span disciplines with different data types, vocabularies, collection methods, and necessary metadata. For example, the descriptive metadata necessary for data collected using *in situ* sensors can be much different than that needed for data collected by *ex-situ* analyses of field specimens. Indeed, one of the major goals in the development of ODM2 was to increase the interoperability of *in situ* and *ex situ* data (Cox 2007a; 2007b) for synthesis analyses, and this drove our selection of data use cases. Each case describes its science driver, data types, benefits from adopting the ODM2 data model, and challenges adapting to the ODM2 data model.

Hydrologic data use case: Little Bear River

The Little Bear River data were collected as part of a National Science Foundation-supported environmental observatory test bed project. The Little Bear River is located in northern Utah, USA, and the project had two main focus areas for data collection: 1) to investigate the use of *in situ* sensor measurements as surrogates for water quality constituents that could not easily be measured continuously (e.g., turbidity as a surrogate for total suspended solids and total phosphorus concentrations) (Spackman Jones et al., 2011; 2012); and 2) to advance the cyberinfrastructure required for collecting, managing, analyzing, and eventually sharing the collected data (Horsburgh et al., 2010a; 2011). The Little Bear River dataset served as a driving use case for the development of the first versions of ODM (Horsburgh et al., 2008) and several components of the HydroServer software stack (Horsburgh et al., 2010b) within the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) Hydrologic Information System (HIS).

The Little Bear River dataset consists of time series of hydrologic observations for over 50 different variables measured using multiple sensors at fixed monitoring sites within the Little Bear River watershed, along with analytical results from water quality samples. Fixed monitoring sites included aquatic sites, with sensors installed within the stream and at which water quality samples were collected, and weather stations at which a standard suite of weather and soil variables were measured. Data collection began in 2005, and is still ongoing at the time of this writing. Water quality samples were collected and analyzed as part of baseline, bi-weekly sampling, along with weekly sampling during spring snowmelt runoff. Additional, event-based samples were collected during rainstorms and significant snowmelt events to capture a wide range of hydrologic conditions within the watershed.

The primary use case of developing surrogate relationships between measurements from *in situ* sensors (e.g., turbidity) and results from water quality samples (e.g., total suspended solids and total phosphorus

concentrations) required integration of both time series from sensors and analytical measurement results derived from water quality samples. The goal was to generate continuous time series of estimated total suspended solids and total phosphorus concentrations by applying a regression model to the *in situ* sensor measurements.

All of the data collected in the Little Bear River were originally integrated into a single ODM 1.1.1 relational database. The database was implemented in the Microsoft SQL Server relational database management system using software tools available from the CUAHSI HIS. This data use case is representative of many common hydrologic studies where observations generated using *in situ* sensors are combined with those derived from specimens collected in the field.

Benefits of using ODM2

The CUAHSI HIS tools were developed with a focus on data publication/sharing. Adoption of ODM 1.1.1 and the additional tools available via the HydroServer software stack capitalized on the data management and sharing capabilities of the CUAHSI HIS. However, while ODM 1.1.1 and associated HydroServer tools met the project's initial needs, several years of data collection highlighted additional data management needs that were not anticipated during the original setup of the Little Bear River monitoring network and in the initial designs of the CUAHSI HIS software.

First, the description of samples, or "Specimens", and their associated collection and analysis procedures was incomplete in ODM 1.1.1. This was a relatively small problem until the types of specimen collections (e.g., grab samples versus automated samples) increased and the number of collected specimens grew into the thousands. More descriptive metadata were needed for each specimen to describe how each was collected, prepared, and analyzed to arrive at the measurement results entered into the database. ODM2 provides a much more extensive ability to not only describe specimens, but also to associate them with actions related to their collection, preparation, and ultimate analysis (**Figure 2**).

Second, over time the sensors and other equipment installed at monitoring sites were subject to maintenance, began to fail, and, in some cases, needed replacement. To ensure the quality and consistency of the data collected at each site, the ability to track which sensors and equipment had been installed at each site over time became an important need. Closely tracking sensor calibrations and environmental conditions affecting sensor measurements to aid in post processing and performing quality control of the *in situ*

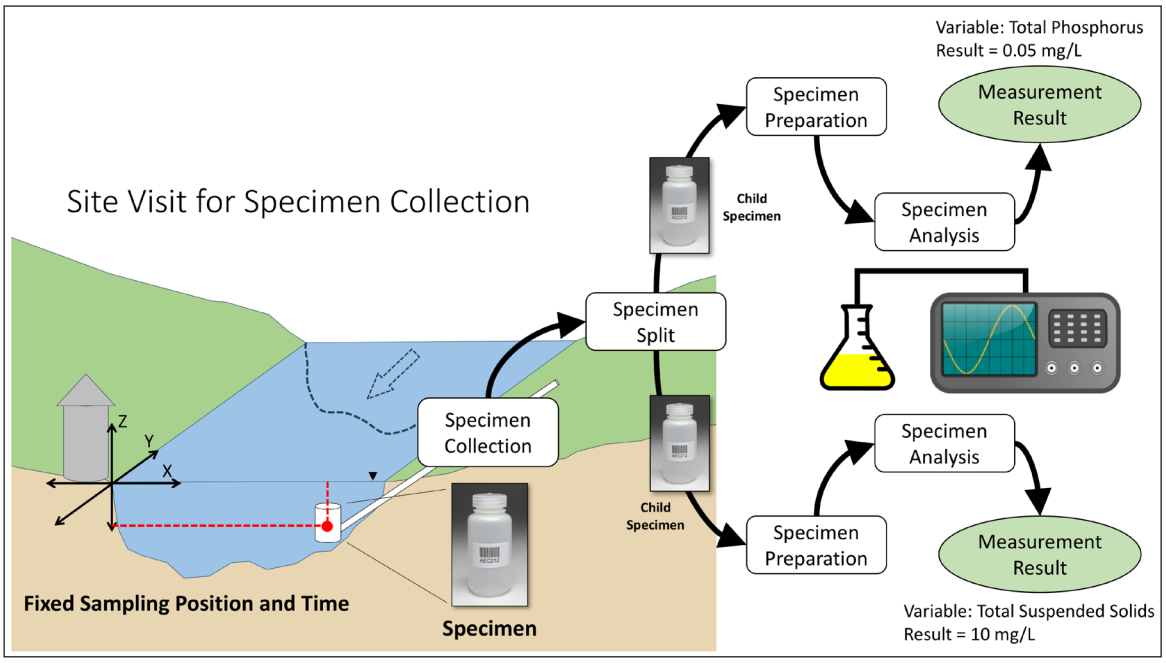

Figure 2: Depiction of a water quality sampling workflow. The separation of sampling features, actions, and results in ODM2 enables a much richer metadata description of specimens and the relationships between them (e.g., parent – child), their collection, preparation, and analysis actions (shown as white boxes in the figure), and the final measurement results.

data was also needed. This level of metadata about field activities, deployed equipment, and environmental observations was not captured in ODM 1.1.1. The Equipment extension in ODM2 provides a place to store and manage much more extensive metadata about sensor and equipment deployments, calibrations, and the field activities surrounding data collection with environmental sensors.

Challenges in adopting ODM2

The process of adapting the Little Bear River ODM 1.1.1 database to a new ODM2 database consisted of creating a structured query language (SQL) script that automatically copies data from the ODM 1.1.1 database to the ODM2 database. The goal in developing this script was to enable anyone who is currently using ODM 1.1.1 to migrate to ODM2. The script contains the mapping for information stored in the ODM 1.1.1 schema to the entities and attributes in the ODM2 schema and is available in the ODM2 GitHub source code repository (https://github.com/ODM2/ODM2/).

ODM2 adopted and built upon most of the concepts from ODM 1.1.1, including Sites, Variables, Methods, Units, QualityControlLevels (now Processing Levels in ODM2), etc. For the Little Bear River use case, mapping these concepts from ODM 1.1.1 to ODM2 was relatively straightforward. However, it was challenging to map or create information that exists in the ODM2 schema (and may be required), but that does not have corresponding content within ODM 1.1.1. For example, ODM 1.1.1 stored information about the "Source" of observations in a single Sources table. ODM2 explicitly models People, Organizations, and Affiliations (a Person's Affiliation with an Organization). While the information from the Sources table in the Little Bear River ODM 1.1.1 satisfies most of what is required for these three entities in ODM2, many of the optional attributes within the three ODM2 entities do not get set by the SQL script because they do not exist as attributes in ODM 1.1.1.

Another challenge was mapping ODM 1.1.1's concept of a "Time Series" to ODM2's concept of a "Result". This was relatively straightforward given that "Time Series" is one of ODM2's Result types. However, ODM2 separates Results from the Actions that created them, and this separation did not exist in ODM 1.1.1. So, moving time series data from an ODM 1.1.1 database to an ODM2 database requires creation of both an Action and a Result, where the Action reflects how the Result was generated via a Method. As a first pass in moving the Little Bear River data, generic "Observation" Actions were created for each time series Result because the ODM 1.1.1 database does not contain information about the sensor deployment Actions. Information about the deployment actions exists in field notes; however, this information would have to be added to the ODM2 database after moving the data across.

Finally, ODM2 adopted most, but not all, of the terms from the ODM 1.1.1 controlled vocabularies (CVs). The ODM2 CVs are more numerous and extensive than the CVs in ODM 1.1.1, and moving the Little Bear River dataset from ODM 1.1.1 to ODM2 required a small amount of semantic moderation prior to running the translation script. This work consisted of checking the CV terms used in the Little Bear River ODM 1.1.1 database for consistency with those terms allowed for ODM2 and making minor modifications. To date, there is no automated tool for performing this work. Although only a small number of inconsistencies were found and fixed within the Little Bear River database, other ODM 1.1.1 users who have made extensive use of ODM 1.1.1 CVs may find this step to be more difficult.

Rock Geochemistry use case: the Petrological Database (PetDB)

PetDB (http://www.earthchem.org/petdb) is a global synthesis of chemical, isotopic, and mineralogical data for rocks, minerals, and melt inclusions, whose current content focuses on data for igneous and metamorphic rocks from the ocean floor, specifically mid-ocean ridge basalts and abyssal peridotites, and xenolith samples from the Earth's mantle and lower crust. One of PetDB's strengths is the compilation of thousands of disparate data publications with millions of analytical values into a synthesized dataset that can be searched based on a number of categorical, geospatial, and qualitative parameters. The synthesis is continuously growing.

The data in PetDB are specimen-based, meaning that the measurements have been made on discrete physical specimens taken from the features being studied. Common specimen types include rocks dredged from the ocean floors, cores drilled from the ocean floor, and rocks manually sampled from on-land outcrops. Measurements made on these whole rock specimens, as well as the volcanic glasses, minerals, and inclusions derived from them include major oxides, trace elements, radiogenic and stable isotopes, analytical age determinations, and more. A wide range of metadata describe specimens (e.g., rock type, texture, age, modal composition, alteration), specimen locations (e.g., geospatial coordinates, location names, tectonic setting), sampling process (e.g., technique, date, cruise), archive, analytical procedures (e.g., method,

precision, standard measurements), and the source of the data (e.g., reference, author(s)). These metadata are not only essential for selecting, sorting, and reusing data properly, but they are fundamental for current and future integration with other data types and interoperability with other databases.

PetDB provides enormous numbers of dense, statistically significant measurements that have driven large, global-scale scientific discoveries. Examples include studies on diversity in mid-ocean ridge basalt (MORB) composition (e.g., Gale et al., 2013) and global patterns of intraplate volcanism (Conrad et al., 2011). PetDB is actively used by the scientific community and has over 600 citations in the literature since 2000 (http://www.earthchem.org/citations/petdb, accessed 3/20/2016). The data model for PetDB was originally developed specifically for mid-ocean ridge basalts (Lehnert et al., 2000). However, a decision was recently made to expand the scope of PetDB to include new data types, in part because of requests by the producers of these data who would like to integrate their data with PetDB. These new data types do not fit well into the original PetDB schema, driving the need for ODM2. Once the new data types are integrated, PetDB will no longer be only for petrologic data, and a name change to EarthChemDB is planned.

Benefits from ODM2

Use of ODM2 for PetDB has allowed integration of data from more specimen types, including soil and volcanic gas specimens. ODM2 enabled PetDB to be more responsive to investigator requests and has enabled PetDB to become a home for disparate data types that may not have other appropriate repositories or that would not be well-curated elsewhere. Furthermore, migrating PetDB to ODM2 has addressed many outstanding needs that were not met with PetDB's original data model, such as the ability to cite the source of metadata (not just the source of the data), allowing multiple locations for a site, the use of external identifiers for people, specimens, and other objects from other established systems, support for time series data, and the ability to better capture parent-child relationships among specimens consisting of rocks, minerals, and inclusions.

Use of ODM2 has allowed unambiguous designation of relationships between sampling features. For example, rocks are composed of minerals, which may have inclusions in them. In the original PetDB, hierarchical relationships between these sampling features were indicated by a careful naming system involving (for marine samples) the cruise, leg, section, and core, so that users and database managers could infer parent-child relationships. In the ODM2 schema, it is possible to define specific relationships between sampling features, thus allowing users to see all mineral or inclusion measurements related to a specific rock, without relying on an elaborate naming scheme to infer relationships (**Figure 3**).

Marine stations (sites) reported in PetDB often have location data that may vary across different publications. This occurs because sites can be revisited, or location data reported in multiple papers may have different positional accuracy. The way the legacy PetDB database dealt with this was to retain all locations that were reported in published articles with the site, but there was no way to record which reference matched with a particular location. In ODM2, the schema allows for multiple locations for the same site, by recording the location (or depth) as an observation made through the Action of a Navigation Measurement. In this way, multiple locations (and the provenance of each) can be stored for a single site.

Challenges in adopting ODM2

PetDB had a number of tables that stored domain-specific data and metadata, such as alteration of a rock sample and heating temperature of an analysis. Because ODM2 is a general schema, those specific fields are no longer included as table columns, but instead are now handled using annotations, for which ODM2 has a general model. The large number of fields that became annotations, and the construction of queries that would efficiently filter by these annotations, was a challenge in moving to the ODM2 schema.

Certain sampling features in PetDB had geometries that were complicated to capture with the ODM2 schema. For example, segments of cores drilled into the ocean floor are commonly identified by their minimum and maximum depth below the ocean bottom. The specialized PetDB schema had fields for minimum and maximum depth to fully describe the geometry. In ODM2, the Spatial Offset could handle this, but not in as simple a way. ODM2 uses three coordinates to describe the offset from an initial point, such as the top of the core. Two values are needed to capture the maximum and minimum, but the concept of maximum and minimum depth is not specifically built into ODM2.

Another challenge was dealing with ODM2 mandatory fields that are not relevant for PetDB data. Certain fields that are necessary for describing sensor measurements are irrelevant or unknown for specimen-based measurements, especially when the only information available is from published literature. The ODM2 concept of an Action is used to capture information about observations performed by a person at a specific time.

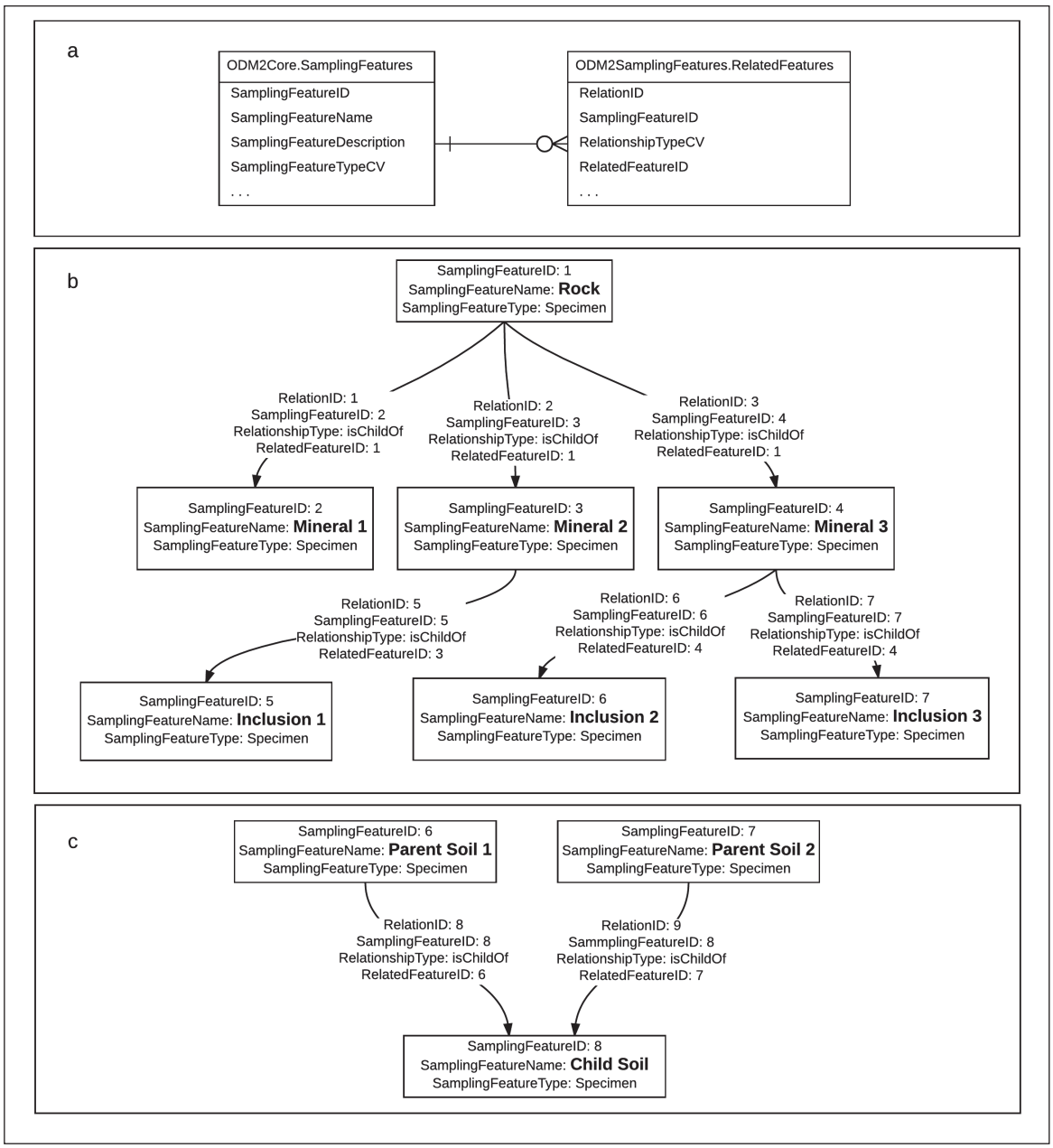

Figure 3: Examples of relationships between parent and child SamplingFeatures in ODM2. **(a)** shows the ODM2 SamplingFeatures and RelatedFeatures entities and their relationship. **(b)** the hierarchy of a rock specimen composed of minerals, which may have inclusions, as an example of multiple child SamplingFeatures derived from a single parent SamplingFeature. **(c)** an example of a soil specimen SamplingFeature created by mixing two parent soil specimen SamplingFeatures. In (b) and (c), SamplingFeatures are shown as boxes, and relationships between them are shown as arrows with text labels showing important attributes of each relationship.

In PetDB, information about people (operators of the analytical geochemical instruments) and time of measurement are usually not provided in the published literature. Therefore, some mandatory fields that are easy to fill out for ongoing lab activities are impossible to fill for PetDB, and default values such as "unknown" were used instead.

Modifications to ODM2

In order to preserve some of the existing functionality of PetDB, some modifications were made to the ODM2 data model. Users of PetDB can search particular geographic locations such as a segment of a mid-ocean ridge, a volcano, or a craton. In the O&M specification on which ODM2 is based (Cox, 2011), these are called "Features of Interest", defined as "a feature that carries the property which is observed". Standard

polygon map searches may not be adequate to outline a Feature of Interest and return all relevant results because boundaries can be fuzzy or interpretive. A simpler approach is to allow users to select from a predefined menu of Features of Interest. Though considered in design discussions, ODM2 currently does not include an entity for describing the Feature of Interest, but the PetDB instance does have one to facilitate searches on physiographic features and tectonic settings.

Another modification has to do with domain-specific data quality information that was captured in PetDB. Data quality information on analytical geochemistry measurements allows users to determine the reliability and comparability of data from different laboratories. There are three concepts for data quality information that are captured in PetDB: fractionation correction, normalization, and standards. To capture the multiple types of data quality information used in PetDB, each reference material is a sampling feature with a description of "reference material". ODM2 was modified to include additional relationship types to relate the analytical results to the data quality values, for example, IsStandardizedBy and IsNormalizedBy.

Soil Geochemistry use case: CZChemDB

Regolith and soil geochemical data provide important observations for understanding the Critical Zone (CZ), the layer of Earth between the subsurface groundwater and the top of the tree canopy (e.g., Banwart et al., 2013). These observations are used to understand processes such as formation of regolith and circulation of fresh water, processes that can greatly affect humans and ecosystems. The Critical Zone Observatory (CZO) program of the US National Science Foundation currently consists of several distinct but connected observatories conducting research on biological, chemical, and physical CZ processes. A shared data infrastructure has been under development since 2013 to promote more efficient scientific understanding across CZO sites (e.g., Zaslavsky et al., 2011), and was one of the main drivers for developing the ODM2 data model.

CZChemDB is a relational database developed to capture geochemical data and relevant documentation in a standardized format to facilitate data sharing among the CZOs (Niu et al., 2011; 2014). The compilation of soil geochemistry in CZChemDB allows both cross-CZO comparisons of observations, and substantially improves data discovery and reuse within a single CZO, where dozens of investigators and students may have generated data over time (e.g., Brantley et al., 2016).

The CZChemDB data model includes metadata describing sampling location, sampling sites, specimens, subspecimens, preparation/treatments, lab analyses, and analytical data values. Supporting descriptive metadata include information about sampling and analytical methods, data quality, data sources, and contributors. The database was originally developed in Microsoft Access for convenience and accessibility. Examples of observed variables include major and trace element and isotopic measurements from samples derived from soil cores and soil pits (e.g., Dere et al., 2013). In addition, these datasets may include physical measurements of particle size, bulk density, and soil color. In weathering studies or studies that examine atmospheric inputs of chemical constituents (e.g., metals derived from industrial activities), chemical and physical measurements of the parent material (e.g., a specimen derived from the underlying parent bedrock) must also be characterized in order to quantify fluxes into or out of the system.

Benefits from ODM2

CZChemDB is an example of a database that was successfully used within a relatively small user base, but required improvements for accessibility, scalability, and interoperability. Adoption of ODM2 addressed each of these needs. For example, moving CZChemDB into a server-based PostgreSQL instance of ODM2 frees it from proprietary, single-user software and enables the data to be more easily accessed by multiple users through client software. Also, the ODM2 information model has stricter constraints than the original Access database, including primary keys and unique constraints on certain entities and fields, which improves integrity of stored and imported data. All of these improvements make the CZChemDB data more robust.

Currently, CZChemDB contains only soil geochemistry. However, investigators in CZOs study much more than soil; they may study everything from the underlying bedrock to the leaves and sap of trees to the porewater, groundwater, surface water, and more. The ability to consistently describe and integrate additional types of data, both specimen- and sensor-based, is therefore critical to allow researchers to form a more complete picture of important processes in order to better understand the fluxes into and out of the CZ (e.g., Brantley et al., 2016). Moreover, integrating multiple data types from multiple CZOs, which may differ climatologically and physiographically, can be facilitated using ODM2 and will allow researchers to investigate and compare the effects of many variables within and across CZOs. It will also allow CZChemDB data to be

integrated with wider geochemical databases such as EarthChemDB that may contain relevant other data, such as compositions of the underlying bedrock.

Like PetDB, hierarchical relationships are very common between specimens in CZChemDB. For example, CZChemDB may contain data for soil samples derived from mixing multiple soil samples together. In this case, the aggregate soil sample is the child of multiple parent soil samples. ODM2 is flexible enough to accommodate complicated relationships between specimens such as this (**Figure 3c**).

Challenges in adopting ODM2

Because CZChemDB focuses on specimen-based geochemical analyses, many of its challenges parallel those in the previous PetDB section, such as relegation of many domain-specific metadata to the Annotation extension. The same modifications to ODM2 made in the PetDB use case were mirrored in CZChemDB, alleviating some of the challenges with data discovery and quality information. Another consequence of being a domain-specific database is that vocabularies and certain phrases were used in ways that are clear to the user base, but not to those in other fields (e.g., the use of the term "CorePitWell" to group those similar entities together). The alignment of vocabularies from CZChemDB to those of the other use cases in ODM2 was challenging.

Initial migration of the CZChemDB data to the ODM2 schema required some additional information. For example, some observations in CZChemDB, such as mean annual precipitation, mean annual temperature, exposure age, erosion rate, and soil taxonomy existed without a citation because the source was assumed or well-known in the community. With the move to a more general model that may be used by non-specialists, ODM2 requires a source for this type of information. When migrating existing CZChemDB data to ODM2, the source was obtained by asking the original investigators when possible, otherwise it was left as "unknown".

Finally, there is a need for specifying public and private data in CZChemDB, since the database is used to keep track of recently collected data that is still in an embargo period. CZChemDB previously accommodated this by maintaining two versions of the Microsoft Access database, one with only the public data, and one with all data. Public and private data are not currently specified in ODM2 or its tools, although a similar, two database solution could be used.

Biogeochemical Time Series Use Case: Marchantaria Amazon River mainstem site

The Marchantaria Time Series is a published (Devol et al., 1995; Richey et al., 1990, 2008), 10-year time series of biogeochemical and basic hydrological measurements in the Amazon River mainstem in central Brazil collected by the Carbon in the Amazon River Experiment (CAMREX) project (Richey et al., 1990; Devol and Hedges, 2001). The data were collected to quantify seasonal and sub-seasonal biogeochemical variability in the mainstem river, and their drivers, for both dissolved and particulate constituents. Samples were collected at intervals that varied but are roughly monthly.

All biogeochemical measurements are based on composite, discharge-weighted river cross-section collection schemes intended to be representative of this wide (>1 km) and deep (up to 30 m) river reach. The sampling scheme used a collapsible bag sampler that filled in proportion to current velocity and was lowered to the bottom and returned at constant speed together with a current meter at multiple points along a channel cross section, while maintaining near-fixed boat positions; individual water samples were then pooled into a single composite sample. Specimens were collected during two types of campaigns: comprehensive cross-section composite sampling during 12 longitudinal cruises along a ~2,000 km length of the mainstem (Richey et al., 1990, 2008); and more frequent, dedicated site visits (101) using a reduced cross-section sample compositing method (Devol et al., 1995). Over 30 biogeochemical and hydrological variables were measured or calculated, addressing several physical fractions (dissolved, fine particulates, coarse particulates, and "total" or bulk composition), plus river stage and estimated discharge and water slope. While this is effectively a "time series" of measurements, it is not based on automated, regular *in situ* sensor observations. For each observed variable, there is an irregular time series of individual measurements, each with corresponding individual result values.

The data was already published, primarily in the supplementary table in Devol et al. (1995) and in the Large-Scale Biosphere-Atmosphere Experiment in Amazonia (LBA)/CAMREX dataset (Richey et al., 2008). It was also previously incorporated in an unpublished, simple relational database created by A. Aufdenkampe and E. Mayorga, originally in Microsoft Access but later migrated to PostgreSQL before mapping to ODM2. Ultimately, this use case is representative of a common situation involving data loading and rescue from old, simple spreadsheets, followed by the use of a simple database with limited scalability.

Benefits from using ODM2

Until now this dataset has been stored in a simple relational database with limited scalability with respect to adding more observed variables and data from other sources and being able to unambiguously distinguish variables by well-defined size fractions. In addition, query and analysis software developed to interact with this database could not be leveraged for wider applications. The database is used only among close collaborators. Sharing the database has carried risks of miscommunication and misuse because methods were not described explicitly and physical fraction distinctions were not well documented, consistent, or generalizable, requiring substantial ancillary information and reference to publications and team knowledge to keep track of these important details. Migrating the data to ODM2 addressed all these limitations, enabling more thorough and granular documentation and provenance, and flexible definitions of size fractions.

ODM2's ability to define relationships among sampling features enables a convenient aggregation of sites into a single "station" (loosely defined) for general analysis. For example, the Marchantaria time series as used in Devol et al. (1995) is actually a combination of semi-regular and targeted sampling at the Marchantaria site, and specimens located in the nearby Manacapuru site. The distinct locations of those two sites can be preserved, while defining a single station to cover both. ODM2's flexibility to retain distinct site provenance while easily enabling the creation of an aggregated station (as a related feature) greatly facilitates analyses that are both robust and convenient.

Porting this dataset to ODM2 facilitated the consistent incorporation of other river biogeochemical data from the Amazon basin, and ultimately from rivers worldwide, without constraints due to a simple data structure. The source database encompasses a large collection of georeferenced biogeochemical measurements across the Amazon River system from both CAMREX and other projects and publications. It has supported the analysis presented in several publications, e.g., Mayorga et al., (2005). Mapping the Marchantaria time series to ODM2 represents the first step in ultimately mapping the entire Amazon database and subsequently making it available via an online application using an ODM2 database backend. Choices made during the Marchantaria data mapping were informed by that longer term goal.

Challenges in adopting ODM2

This dataset was collected to be used as a time series. Collection and analysis methods remained fairly consistent, though there was some variability. Sampling frequency was quite variable due to logistical challenges. However, in the context of ODM2 Result Types, time series (more specifically, "Time Series Coverage") has the specific meaning of being a highly regular data collection resulting from identical methods, typically from automated, *in situ* sensor observations. Identifying a dataset of measurements as effectively a "time series" (though a sparse and irregular one) required developing certain conventions for querying and categorization. Thus, this use case relies on Specimen Sampling Features and Measurement Results. Time series composed of analyses of samples collected at semi-regular intervals (e.g., weekly or monthly) and using methods that may vary over the years are quite common in water quality monitoring, and in biogeochemical and other earth sciences research, making this a challenge to many applications.

The database where these data were held had severe limitations in its ability to represent analytical and sampling methods. This information had to be added in the code used to map the old database to ODM2, and often it was only available in the publication's Methods section. This issue does not reflect a shortcoming of ODM2 proper, but rather the difficulty of mapping from simpler data storage schemes to take advantage of ODM2 capabilities for capturing more resolved information. Using methods that represent an aggregated series of steps was often necessary, but may result in degraded interoperability with ODM2 databases with more resolved and specific methods.

ODM2 often provides more than one way to implement a concept or relationship. When dealing with data extracted from the literature, the simplifications that are involved (such as description of methods) can make those decisions less clear. For example, at this time information about the sample physical fraction (e.g., fine versus coarse suspended particulates) was placed in both SamplingFeatures via the definition of specimens, and in the Variables definition via the use of highly specific particulate fraction variable names. Some variables were created that are specific to each fraction (e.g., FSS versus CSS for Fine and Coarse Suspended Sediment, respectively). The use of such variables together with specimens by physical fraction is somewhat redundant, but convenient. The development of a best practice for handling these physical fraction distinctions in ODM2 would greatly facilitate queries across databases and possibly across domains. A similar challenge and need for documented best practices arose in deciding whether a specimen sampling feature should include the geospatial coordinates, or whether this geospatial information should only be encoded in the site sampling feature linked via related features.

Lessons learned from ODM2 implementation

Similarities across use cases

The use cases and their implementations are summarized in **Tables 1** and **2**. The benefits of ODM2 are diverse and vary with each use case, but there are a few common themes. The previous data models did not accommodate all of the use case systems' needs, which typically happened because many of the data systems were growing their user base and adding new requirements. The more general ODM2 information model provided solutions to these new requirements, but, at the same time, often reduced some of the convenience and efficiency of the more specific data models. In other words, ODM2 made the systems much more extensible, which almost all use cases needed, at the expense of being less optimized for specific data types.

The sampling feature concept, which combines specimens and sites, and the way that geometries are specified, benefited most cases. Specifically, the flexibility to define different relationships between sampling features, whether hierarchically or in groups, was a common benefit.

ODM2's general ability to capture metadata for collecting, managing and analyzing samples or measurements was another common benefit. Metadata for different use cases are complex and varied, and were the root of many of the challenges met in implementation. The specialized information that is critical for assessing data quality and reuse for other scientific domains should not be underestimated. The line between data and metadata can be unclear, especially when considering multiple use cases. Information that might be critical for assessing the usability of data in one domain might seem trivial or unimportant in others. The distinction between data and metadata may not need to be pinned down as long as both are available for use in applications (Mookerjee et al., 2015). A major advantage of ODM2 is that it encourages consistent description of concepts that are common across data types (e.g., variables, units, sampling features, etc.) while allowing more specific metadata for specialized data types (e.g., *ex situ* measurements versus *in situ* time series). More detail on the specific use case implementations can be found in the ODM2 GitHub repository at (https://github.com/ODM2/ODM2/tree/master/usecases).

Name	Hydrology: Little Bear River	Rock Geochemistry: PetDB	Soil Geochemistry: CZChemDB	Biogeochemistry: Marchantaria
Organization	Utah State University, Utah Water Research Laboratory	IEDA, Interdisciplinary Earth Data Alliance, Lamont Doherty Earth Observatory, Columbia University	Susquehanna Shale Hills Critical Zone Observatory, Pennsylvania State University	CAMREX, Carbon in the Amazon River Experiment (University of Washington and CENA/Universidade de Sao Paulo, Brazil)
Domain focus	hydrology, water quality	geochemistry, rocks, minerals, and inclusions	geochemistry, soils and regolith	aquatic geochemistry biogeochemistry, hydrology
Specimen, Time Series, or both?	both specimens and time series	specimen	specimen	specimen
Primary data management and generation, or synthesis/ literature?	Data management for a research watershed. Data generated by in situ sensors and regular and event based field sampling of water quality.	synthesis of published literature	synthesis and data management	synthesis of data generated by the research project, but compiled only after publication
Time period	2005 – Present. Some data collection ongoing.	earliest publication is 1937 to present (ongoing)	2010 to present	1982–1993
Spatial domain	Little Bear River Watershed, Utah, USA	Global	Shale Hills Critical Zone Observatory, Other Critical Zone Observatories	Central Amazon mainstem river site, Brazil

Table 1: ODM2 Use Cases.

Name	Hydrology: Little Bear River	Rock Geochemistry: PetDB	Soil Geochemistry: CZChemDB	Biogeochemistry: Marchantaria
Changes/additions to ODM2 schema?	no	yes (see text)	yes (see text)	no
Other, external CVs?	no	yes	yes	no
Original RDBMS system used[1]	Microsoft SQL Server	Oracle	Microsoft Access	Microsoft Access and PostgreSQL
RDBMS implemented for ODM2	Microsoft SQL Server	PostgreSQL	PostgreSQL	PostgreSQL
# Sampling Feature Sites[2]	16	22,800+ "stations"	265+	2 "water quality stations"
# Sampling Feature Specimens[2]	3600+	83,000+	1980+	700+
# Results and Results Values[2]	~550 results, 30 million values	3,000,000+ values	22,700+ values	2,510+ values
Result Types used	Time Series, Measurement	Measurement	Measurement	Measurement

Table 2: ODM2 Implementation Summary.

[1] For data models used in the original data sources, see the Use Cases section.

[2] Data collection and/or data mapping and loading are ongoing in all use cases. These values represent the state of the ODM2 mappings as of March 2016.

Common challenges for ODM2 adopters

Almost all systems and people that contributed to the design of ODM2 agreed that controlled vocabularies were a benefit to interoperability, but achieving agreement on the vocabularies was difficult. ODM2 encourages use of standard vocabularies (http://vocabulary.odm2.org) but acknowledges that the legacy of production systems and ingrained domain culture make it difficult to fully accept standard vocabularies without modification. Thus, ODM2 allows for using terms from any formally published vocabularies. There are, however, many vocabularies in use that are not formally published, and so a push to get community-supported vocabularies published and guidance on how to use the available options would be beneficial. In addition, attempts to generate thesauri and other semantic mappings would help the effort for interoperability.

Once a flexible and general information model is agreed on by different disciplines, templates for data entry and tools for data ingestion are an entirely separate challenge. Initial data migration may be completed by scripts, but after that, the problem of new data ingestion remains. Generality in a data model can make terms in the schema more ambiguous and less recognizable for domain researchers and data managers (e.g., use of the term "sampling feature" to describe both sites and specimens). Also, researchers used to simple spreadsheet formats with minimal metadata may need to deal with increasingly unwieldy spreadsheets containing more complicated metadata. The challenge of data entry, particularly for specimen observations, is being addressed by an effort by the ODM2 team to create Microsoft Excel-based data entry templates that can be exported into a text file format that is both human and machine readable for ingestion into an ODM2 database instance or for transfer over the internet (Horsburgh et al., 2016).

Best Practices for ODM2 Implementation

Our experiences with data use cases and adoption of ODM2 have highlighted some best practices for promoting interoperability of systems when adopting a new general information model.

- **Use external identifiers and external registries where possible – e.g., for People, Institutions, Citations, Methods, Samples.** To decrease redundancy, duplication, and variation, use existing authoritative or community-supported registries. For example, use ORCiDs for people, International Standard Name Identifiers (ISNIs) for institutions, digital object identifiers

(DOIs) for Citations, National Environmental Methods Index (NEMI) identifiers for Methods, and International Geo Sample Numbers (IGSNs) for specimens (e.g., Hanson, 2016). Using formal identifiers ensures that linkages are maintained between instances of information stored in ODM2 databases and the content in its original or authoritative source and that updates can be more easily made when needed.

- **Consider using ODM2's recommended controlled vocabularies, or start from them.** As an additional measure to decrease redundancy, duplication, and unnecessary semantic variability, consider using ODM2's CVs (http://vocabulary.odm2.org). This system enables contributions of needed terms from the community, which means that it can adapt through use. It is easier to accommodate cases not considered in the recommended vocabularies in order to make them more robust for more users, rather than retroactively trying to map a new vocabulary. ODM2's vocabularies were developed and adapted from multiple sources, including the CUAHSI HIS and ODM 1.1.1, the United States Geological Survey's National Water Information System, and IEDA's PetDB and other systems. Thus, they cover a broad range of potential data use cases from continuous hydrologic sensor time series to solid earth geochemical samples and can be extended as needed. As described above, in the Little Bear River use case a small amount of semantic mediation was required to ensure that vocabulary terms used in the ODM 1.1.1 database were compatible with the ODM2 vocabularies. Where possible, it is recommended that any mediation work be done before trying to move data into an ODM2 instance to minimize potential errors and inconsistencies that might be introduced.
- **Use Annotations and/or the ODM2 Provenance extension to document provenance of information.** The ODM2 Annotations extension allows qualifying comments or notes for Sampling Features, Actions, Methods, Results, Result Values, and Equipment. With minor modification, annotations can be applied to other ODM2 entities. Don't assume that other users of your data will have the same disciplinary common knowledge that you have; interdisciplinary collaborations will benefit from provenance and sources. In the PetDB use case, the source of data is peer reviewed literature, and each publication contains a vast store of contextual information. The ability to include this contextual data, while still giving credit to the source of the contextual data was extremely important, and made possible by using the Annotations and Provenance extensions.
- **Write scripts for the initial migration of data.** In addition to serving as a reproducible record of migration, the scripts may inform other adopters on how to migrate data to the new data model. Reproducible migration was necessary in the PetDB use case, which involves a widely-used operational system with frequent data entry. Scripts documenting reproducible procedures allowed PetDB to continue to add new data as migration was ongoing and new data entry methods were being developed for the new data model. Similarly, scripting was used to migrate the Little Bear River ODM 1.1.1 database to an instance of ODM2, and the resulting SQL script now serves as an example of how other research groups who have used ODM 1.1.1 can map and migrate their data to ODM2.
- **Have a primary data collector advise on the data migration and requirements.** We found that adoption of the data model was many times easier when a producer of data like those being migrated was available to explain the details of how the data and metadata would be used. For specimen-based data, understanding of how samples were collected and, most importantly, the relationship between samples and subsamples, was essential for making use of the data model. For sensor-based observations, it was helpful to understand how sensors are deployed and maintained to better enable capturing field deployment, calibration, and maintenance actions that can be important in interpreting data but that are rarely recorded with data and reported.

Summary

The development of the ODM2 information model was motivated by the practical challenge of integrating spatially discrete Earth observations from diverse domains and data types. As Earth science problems become more interdisciplinary and teams more collaborative, this issue will come up more frequently. The ODM2 information model was born from discussions of a diverse team composed predominantly of practicing geoscientists who have experience with operational data models that needed improvement. Thus, the design was strongly driven by requirements coming directly from research science needs. The ODM2 effort has added large amounts of implementation detail and practical testing to many core ideas from the original Observations & Measurements data model (Cox, 2007a; 2007b; 2011).

For existing data systems, the motivation for adopting a new, general information model like ODM2 must be that it provides a solution for a challenge that is not being met: there must be clear incentives. ODM2 provides motivation through scalability, interoperability, and related software tools being developed for management, visualization and analysis of data stored in ODM2 database instances (see software development activities in the ODM2 organization in GitHub – http://www.github.com/ODM2/). Additionally, we are now working on new web application software for publishing data stored in an ODM2 database using standards-based web services. New data systems that are looking for a ready-made, interoperable information model are an easier case for ODM2 implementation.

This paper provides a wide-ranging illustration and discussion of the practical challenges (and successes) faced when implementing the ODM2 information model, and more generally when migrating data to broader data models. Though all use cases can list significant benefits from moving to the more general ODM2 model, they all also experienced challenges in handling domain-specific attributes. It is anticipated that an up-front investment in making disparate data systems interoperable through adoption of a common information model will lead to significant returns, including enhanced metadata for data discovery, better accessibility through more robust database instances, interoperability through common data descriptions and vocabularies, and enhanced analyses supported by more descriptive metadata.

Acknowledgements

The authors gratefully acknowledge the contributions of participants at the ODM2 community design workshops and two reviewers. This work was supported by the National Science Foundation under grant EAR 1224638. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Competing Interests

The authors have no competing interests to declare.

Author Information

Leslie Hsu is now at U.S. Geological Survey, Denver, CO 80225, USA.

References

Banwart, S, Chorover, J, Gaillardet, J, Sparks, D, White, T, et al. 2013 Sustaining Earth's Critical Zone – Basic Science and Interdisciplinary Solutions for Global Challenges. The University of Sheffield, United Kingdom, ISBN: 978-0-9576890-0-8.

Brantley, S L, DiBiase, R, Russo, T, Shi, Y, Lin, H, Davis, K J, Kaye, M, Hill, L, Kaye, J, Neal, A L, Eissenstat, D, Hoagland, B and **Dere, A L** 2016 Designing a suite of measurements to understand the critical zone. *Earth Surf. Dynam. Discuss.*, 3: 1005–1059. DOI: https://doi.org/10.5194/esurfd-3-1005-2015

Brantley, S L, Goldhaber, M B and **Ragnarsdottlr, K V** 2007 Crossing disciplines and scales to understand the critical zone. *Elements*, 3: 307–314. DOI: https://doi.org/10.2113/gselements.3.5.307

Conrad, C P, Bianco, T A, Smith, E I and **Wessel, P** 2011 Patterns of intraplate volcanism controlled by asthenospheric shear. *Nature Geosci*, 4: 317–321. DOI: https://doi.org/10.1038/ngeo1111

Cox, S J D 2007a Observations and Measurements – Part 1 – Observation schema v1.0, OGC 07-022r1. 73 + xi. Retrieved from http://portal.opengeospatial.org/files/22466.

Cox, S J D 2007b Observations and Measurements – Part 2 – Sampling Features v1.0, OGC 07-002r3. 36 + ix. Retrieved from http://portal.opengeospatial.org/files/22467.

Cox, S 2011 Geographic Information: Observations and Measurements OGC Abstract Specification Topic 20, v2.0.0. OGC 10-004r3. Open Geospatial Consortium, Inc., p. 49. Retrieved from http://portal.opengeospatial.org/files/?artifact_id=41579 (last accessed 16-04-29).

Dere, A L, White, T S, April, R H, Reynolds, B, MIller, T E, Knapp, E P, McKay, L D and **Brantley, S L** 2013 Climate dependence of feldspar weathering in shale soils along a latitudinal gradient. *Geochimica et Cosmochimica Acta*, 122: 101–126. DOI: https://doi.org/10.1016/j.gca.2013.08.001

Devol, A H, Forsberg, B R, Richey, J E and **Pimentel, T P** 1995 Seasonal variation in chemical distributions in the Amazon (Solimoes) River: A multiyear time series. *Global Biogeochemical Cycles*, 9(3): 307–328. DOI: https://doi.org/10.1029/95GB01145

Devol, A H and **Hedges, J I** 2001 Organic matter and nutrients in the mainstem Amazon River. In: McClain, M E, Victoria, R L and Richey, J E (Eds.) *The biogeochemistry of the Amazon basin*. Oxford University Press, New York, pp. 275–306.

Gale, A, Dalton, C A, Langmuir, C H, Su, Y and **Schilling, J-G** 2013 The mean composition of ocean ridge basalts. *Geochem. Geophys. Geosyst.*, 14: 489–518. DOI: https://doi.org/10.1029/2012GC004334

Hanson, B 2016 AGU opens its journals to author identifiers. *Eos*, 97. DOI: https://doi.org/10.1029/2016EO043183

Horsburgh, J S, Aufdenkampe, A K, Mayorga, E, Lehnert, K A, Hsu, L, Song, L, Spackman Jones, A, Damiano, S G, Tarboton, D G, Valentine, D, Zaslavsky, I and **Whitenack, T** 2016 Observations Data Model 2: A community information model for spatially discrete Earth observations. *Environmental Modelling & Software*, 79: 55–74. DOI: https://doi.org/10.1016/j.envsoft.2016.01.010

Horsburgh, J S, Spackman Jones, A, Tarboton, D G, Stevens, D K and **Mesner, N O** 2010a A sensor network for high frequency estimation of water quality constituent fluxes using surrogates. *Environmental Modelling & Software*, 25: 1031–1044. DOI: https://doi.org/10.1016/j.envsoft.2009.10.012

Horsburgh, J S, Tarboton, D G, Maidment, D R and **Zaslavsky, I** 2008 A relational model for environmental and water resources data. *Water Resour. Res.*, 44: W05406. DOI: https://doi.org/10.1029/2007WR006392

Horsburgh, J S, Tarboton, D G, Maidment, D R and **Zaslavsky, I** 2011 Components of an environmental observatory information system. *Computers & Geoscience*, 37(2): 207–218. DOI: https://doi.org/10.1016/j.cageo.2010.07.003

Horsburgh, J S, Tarboton, D G, Schreuders, K A T, Maidment, D R, Zaslavsky, I and **Valentine, D** 2010b HydroServer: A platform for publishing space-time hydrologic datasets. In: *Proceedings of the AWRA Spring Specialty Conference on GIS and Water Resources*, Orlando, FL, March 29–31.

Lehnert, K, Su, Y, Langmuir, C H, Sarbas, B and **Nohl, U** 2000 A global geochemical database structure for rocks. *Geochem. Geophys. Geosyst.*, v. 1: 1999GC000026. DOI: https://doi.org/10.1029/1999GC000026

Mayorga, E, Aufdenkampe, A K, Masiello, C A, Krusche, A V, Hedges, J I, Quay, P D, Richey, J E and **Brown, T A** 2005 Young organic matter as a source of carbon dioxide outgassing from Amazonian rivers. *Nature*, 436: 538–541. DOI: https://doi.org/10.1038/nature03880

Mookerjee, M, Vieira, D, Chan, M A, Gil, Y, Pavlis, T L, Spear, F S and **Tikoff, B** 2015 Field data management: Integrating cyberscience and geoscience. *Eos*, 96. DOI: https://doi.org/10.1029/2015EO036703

Niu, X, Lehnert, K A, Williams, J and **Brantley, S L** 2011 CZChemDB and EarthChem: Advancing management and access of critical zone geochemical data. *Applied Geochemistry*, 26: S108–S111. DOI: https://doi.org/10.1016/j.apgeochem.2011.03.042

Niu, X, Williams, J Z, Miller, D, Lehnert, K, Bills, B and **Brantley, S L** 2014 An Ontology Driven Relational Geochemical Database for the Earth's Critical Zone: CZchemDB. *Journal of Environmental Informatics*, 23(2): 10–23. DOI: https://doi.org/10.3808/jei.201400266

Richey, J E, Hedges, J I, Devol, A H, Quay, P D, Victoria, R, Martinelli, L and **Forsberg, B R** 1990 Biogeochemistry of carbon in the Amazon River. *Limnology & Oceanography*, 35: 352–371. DOI: https://doi.org/10.4319/lo.1990.35.2.0352

Richey, J E, Victoria, R L, Hedges, J I, Dunne, T, Martinelli, L A, Mertes, L and **Adams, J** 2008 Pre-LBA Carbon in the Amazon River Experiment (CAMREX) Data. Data set., from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A. Available at: http:/daac.ornl.gov. DOI: https://doi.org/10.3334/ORNLDAAC/904

Spackman Jones, A, Horsburgh, J S, Mesner, N O, Ryel, R J and **Stevens, D K** 2012 Influence of sampling frequency on estimation of annual total phosphorus and total suspended solids loads. *Journal of the American Water Resources Association*, 48(6): 1258–1275. DOI: https://doi.org/10.1111/j.1752-1688.2012.00684.x

Spackman Jones, A., Stevens, D K, Horsburgh, J S and **Mesner, N O** 2011 Surrogate measures for providing high frequency estimates of total suspended solids and total phosphorus concentrations. *Journal of the American Water Resources Association*, 47(2): 239–253. DOI: https://doi.org/10.1111/j.1752-1688.2010.00505.x

Wendl, M C, Smith, S, Pohl, C S, Dooling, D J, Chinwalla, A T, Crouse, K, Hepler, T, Leong, S, Carmichael, L, Nhan, M, Oberkfell, B J, Mardis, E R, Hillier, L W and **Wilson, R K** 2007 Design and implementation of a generalized laboratory data model. *BMC Bioinformatics*, 8: 362. DOI: https://doi.org/10.1186/1471-2105-8-362

Zaslavsky, I, Whitenack, T, Williams, M, Tarboton, D G, Schreuders, K and **Aufdenkampe, A** 2011 The Initial Design of Data Sharing Infrastructure for the Critical Zone Observatory. In *Proceedings of the Environmental Information Management Conference, Santa Barbara, CA*, 28–29 September, EIM'2011, pp. 145–150.

How to cite this article: Hsu, L, Mayorga, E, Horburgh, H S, Carter, M R, Lehnert, K A and Brantley, S L 2017 Enhancing Interoperability and Capabilities of Earth Science Data using the Observations Data Model 2 (ODM2). *Data Science Journal,* 16: 4, pp. 1–16, DOI: https://doi.org/10.5334/dsj-2017-004

Submitted: 30 April 2016 **Accepted:** 19 January 2017 **Published:** 06 February 2017

]u[*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS