HydroShare: Sharing Diverse Environmental Data Types and Models as Social Objects with Application to the Hydrology Domain

Jeffery S. Horsburgh [1], Mohamed M. Morsy, Anthony M. Castronova, Jonathan L. Goodall, Tian Gan, Hong Yi, Michael J. Stealey, David G. Tarboton

This is the accepted version of the following article:

Horsburgh, J. S., M. M. Morsy, A. M. Castronova, J. L. Goodall, T. Gan, H. Yi, M. J. Stealey and D. G. Tarboton, (2016), "Hydroshare: Sharing Diverse Environmental Data Types and Models as Social Objects with Application to the Hydrology Domain," JAWRA Journal of the American Water Resources Association, 52(4): 873-889, http://dx.doi.org/10.1111/1752-1688.12363.

which has been published in final form at http://dx.doi.org/10.1111/1752-1688.12363.

[1] Assistant Professor (Horsburgh), Department of Civil and Environmental Engineering and Utah Water Research Laboratory, Utah State University, 8200 Old Main Hill, Logan, UT, 84322-8200, jeff.horsburgh@usu.edu; PhD Candidate (Morsy), Department of Civil and Environmental Engineering, University of Virginia, Charlottesville, VA, 22904; Research Assistant Professor (Castronova), Department of Civil and Environmental Engineering and Utah Water Research Laboratory, Utah State University, Logan, UT, 84322-8200; Associate Professor (Goodall), Department of Civil and Environmental Engineering, University of Virginia, Charlottesville, VA, 22904; PhD Candidate (Gan), Department of Civil and Environmental Engineering and Utah Water Research Laboratory, Utah State University, Logan, UT, 84322-8200; Senior Research Software Developer (Yi), Renaissance Computing Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27517; Senior Research Software Developer (Stealey), Renaissance Computing Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27517; Professor (Tarboton), Department of Civil and Environmental Engineering and Utah Water Research Laboratory, Utah State University, Logan, UT, 84322-4110

ABSTRACT

The types of data and models used within the hydrologic science community are diverse. New repositories have succeeded in making data and models more accessible, but are, in most cases, limited to particular types or classes of data or models and also lack the type of collaborative, and iterative functionality needed to enable shared data collection and modeling workflows. File sharing systems currently used within many scientific communities for private sharing of preliminary and intermediate data and modeling products do not support collaborative data capture, description, visualization, and annotation. In this paper we cast hydrologic datasets and models as “social objects” that can be published, collaborated around, annotated, discovered, and accessed. This paper describes the generic data model and content packaging scheme for diverse hydrologic datasets and models used by a new hydrologic collaborative environment called HydroShare to enable storage, management, sharing, publication, and annotation of the diverse types of data and models used by hydrologic scientists. The flexibility of HydroShare’s data model and packaging scheme is demonstrated using multiple hydrologic data and model use cases that highlight its features.

Key Terms: Data management; Open source software; Hydrologic information systems; HydroShare; Data sharing, Collaborative environment

INTRODUCTION

Emerging data repositories in the geosciences are doing a tremendous job of increasing the availability of environmental datasets and better supporting the long tail of scientific data (Heidorn, 2008). These include the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) Hydrologic Information System (Tarboton et al., 2009), the Critical Zone Observatory Integrated Data Management System (CZOData) (Zaslavsky et al., 2011), the Integrated Earth Data Applications and EarthChem system (Lehnert et al., 2011), the Integrated Ocean Observing System (De La Beaujardiere, 2008), the Data Observation Network for Earth (DataONE) (Michener et al., 2011), among others. However, outside of larger domain cyberinfrastructure development efforts, many datasets are not published, or, if they are shared, it is in primitive formats that are hard to find, difficult to interpret, and do not express the knowledge and insights of the data collector that could be applied to the next study that uses the data. As a result, our current system for publishing scientific knowledge contains only a fraction of the data we collect. Better infrastructure is needed for the full range of scientific activities, including data capture, curation, analysis, and publication.

Data sharing and publication are important in ensuring reproducible science (e.g., Costello and Wieczorek, 2014; HSU et al., 2015). Scientists wish to (or may be required by funding agencies or journals to) publish their data with their results to ensure that others can reproduce their work. Some are even calling for more formal publication and peer review of datasets (e.g., Costello et al., 2013). While peer review of published data is currently uncommon, existing research data infrastructures, including an increasing

number of research libraries (Tenopir et al., 2014), generally support the data publication use case. Scientists can deposit finished results of their analyses into a repository and share them with the world. However, researchers may want to privately share preliminary or unfinished versions of their data products with colleagues or collaboratively iterate through multiple versions of a product and its metadata description prior to publication. An online system for collaboration can facilitate the early capture of data in a platform independent system, and new collaboration functionality can enable products and metadata to evolve before eventual publication. However, file sharing systems like Drop Box or Google Drive, which are commonly used now within many scientific communities for private sharing of preliminary and intermediate data products, do not support this type of collaborative data capture, description, visualization, annotation, etc. Existing data publication repositories do not currently enable this type of collaboration either.

Another challenge is that products deposited into research data infrastructures are generally project/study specific. Potential data users may struggle to determine whether available data is appropriate for a new use. Costello et al. (2013) describe how potential data users must currently study both the metadata and the process used to create the data to establish fitness for a specific purpose, whereas enhancements to metadata would help potential users understand appropriate uses.

Computational models pose a different challenge. For the purposes of this paper, we consider models as being comprised of two distinct resources: 1) the model logic as source code or compiled executable (what we call here a *Model Program*) and 2) the input files for a specific location and time period used to run the model along with the

output generated by the model (what we call here a *Model Instance*). These two resources are interconnected in a one-to-many relationship, wherein one Model Program can be used to execute many Model Instances. This enables Model Instances to be shared (i.e., a particular analysis with all of its inputs and outputs), which is necessary to ensure that study results can be reproduced.

There have been efforts to create general model sharing repositories and metadata standards to support such repositories. For instance, the Community Surface Dynamics Modeling System (CSDMS) project developed a repository (http://csdms.colorado.edu/wiki/Model_download_portal) that provides metadata for a large number of models used in the geosciences. CSDMS and others have proposed standards for model components, model metadata, and controlled vocabularies (Gregersen et al., 2007; Nagai et al., 2012; Elag and Goodall, 2013; Peckham et al., 2013; Peckham, 2014). However, there has been less focus on sharing Model Instances in a formal way that is well documented and associates instances with Model Programs to support reproducible science (e.g., Dunlap et al., 2008). What is needed is a system that can capture the structure and metadata of a Model Programs, Model Instances, and the relationships between them for the purpose of sharing among collaborators.

In this paper we cast hydrologic datasets and models as “social objects” that can be published, collaborated around, annotated, discovered, and accessed. Thus far, domain cyberinfrastructures for data publication have focused on important data classes (e.g., the CUAHSI HIS focused on hydrologic time series datasets). However, the types of data and models used within the hydrologic science community are diverse. We needed

to move beyond time series to better support the types of sharing and collaboration needed in the hydrology community. This paper focuses on the generic data model and content packaging scheme for diverse hydrologic datasets and models that are foundational within a new hydrologic collaborative environment called HydroShare to overcome the challenges described above.

HYDROLOGIC DATA AND MODELS AS SOCIAL OBJECTS

Social objects are objects around which social networks form (Engeström, 2005). For example, many social media websites such as Pinterest (https://www.pinterest.com), Flickr (https://www.flickr.com), YouTube (https://www.youtube.com), and others enable sharing of photographs and videos as social objects that can be viewed, tagged, commented on, and re-shared. Within many scientific disciplines, datasets and models have become social objects around which collaborations and networks form. It is now common for groups of scientists working within interdisciplinary projects to collect shared datasets or develop collaborative models. Examples include the National Science Foundation (NSF) funded Critical Zone Observatories (CZOs) in the U.S. (http://criticalzone.org/national/). Similarly, some scientific communities have models that are accepted by and advanced collaboratively within the community (e.g., the Weather Research and Forecasting (WRF) model (http://www.wrf-model.org) used by communities of atmospheric scientists and climate scientists).

For some of these efforts, like the CZOs, new cyberinfrastructure is emerging for publishing shared datasets on the Internet (Zaslavsky et al., 2011), and the availability

of data across networks of sites is increasing. However, existing or emerging cyberinfrastructures have focused on publication and discovery and not collaborative development or use of the data. They do not yet enable individual users/scientists to easily create digital instances of datasets and models (i.e., the social objects), quickly share them with their colleagues, and add value by annotating them with information about scientific use cases for which they are well suited, inherent limitations, conclusions that have been drawn, or interesting findings. Instead, a relatively small number of data managers and technicians act as curators for finished datasets and research products produced by larger projects or research groups. Outside of national observatory efforts, few tools and repositories are available for smaller research groups to share and collaborate around their data.

HydroShare: A Hydrologic Information System for Sharing Social Objects

HydroShare (http://www.hydroshare.org) is a next-generation, Internet-based hydrologic information system for sharing and collaborating around hydrologic data and models. Tarboton et al. (2014a) provide a broad overview of the functionality envisioned for HydroShare, and Heard et al. (2014) provide a description of the software architecture on which HydroShare is built. The main goal of HydroShare is to facilitate creation, collaboration around, discovery of, and access to data and model resources shared by members of the Hydrology community. HydroShare aims to provide collaborative social functions for datasets and models, including both private and public sharing, formation of collaborative groups, and value-added annotation of digital content. Content that can be shared within HydroShare is diverse, including digital

objects that represent multiple hydrologic data types, models and model instances, documents, and other content types commonly used in hydrologic research. These include hydrologic time series, geographic features (vector data), geographic rasters (gridded data), multidimensional space-time data sets (e.g., NetCDF), and composite resources that represent complex datasets such as river geometry. Model Programs and Model Instances are additional types of content that can be shared and manipulated within HydroShare.

A “resource” is the discrete unit of digital content within HydroShare. Resources are social objects that can be created, modified, versioned, shared, annotated, discovered, and accessed. In this resource-centric approach, which was briefly introduced in Tarboton et al. (2014b), a resource is the granular unit used for management and access control within HydroShare. System metadata is maintained that tracks system-level attributes of the resource, including time stamps of creation and modification, ownership, access control rules, etc. Persistent identifiers, access control, versioning, sharing, and discovery are all managed at the resource level in HydroShare. The following are properties of HydroShare resources:

1. A resource may be made up of a single content file (e.g., in the case of a file containing a single hydrologic time series) or may be an aggregation of multiple content files (e.g., in the case of a hydrologic Model Instance with various input files necessary for its execution).
2. A resource containing multiple content files may have a hierarchical file/directory structure.

3. A resource may conform to a standardized content data model that is specific to a particular resource type and may define specific file formats, syntax, and/or file hierarchies.
4. A resource is described by resource-level metadata that detail its properties. Resource-level metadata may contain extensions that are specific to a particular resource type.
5. Each content file within a resource may be separately described by content-level metadata (e.g., a separate metadata document that describes a specific file or group of files within a resource).

Given the diversity in the types of content supported by HydroShare, we needed to overcome several major technical challenges. First, HydroShare needed to be able to store structured resources within the system so that social metadata such as comments, ratings, and annotations could be associated with resources and so that value added software tools that operate on known resource structures could be built. However, we did not want to impose a single structure that would limit the potential types of resources that can be shared. Next, we didn't want to force users to do extensive reformatting of data prior to uploading it to HydroShare, rather we wanted to support the use of common data formats already used within the hydrologic science community. In fact, we compiled the list of resource properties above by examining the data types and formats commonly used with the hydrology community. We also needed a method for describing and then extracting consistent metadata from all resource types so they could be cataloged for discovery purposes. Finally, HydroShare required a flexible packaging format for consistently storing all types of resources on disk and

packaging them for transport over the Internet (i.e., for download from and upload to HydroShare). In the following sections we describe HydroShare's resource data model and packaging scheme, which were designed to meet these use cases. Following that, we provide some specific examples of how common hydrologic data types are stored and packaged by HydroShare.

A DATA MODEL FOR HYDROLOGIC DATA AND MODEL RESOURCES

HydroShare's overarching resource data model is an implementation of the Open Archives Initiative's Object Reuse and Exchange (OAI-ORE) standard (Lagoze et al., 2008b). OAI-ORE is a standard for the description and exchange of aggregations of web resources, where "aggregations" in this case means related groups of computer files. The OAI-ORE Abstract Data Model, including definitions of the data model entities, is described by Lagoze et al. (2008a) and is well suited for representing the type of file aggregations we needed for HydroShare resources. Figure 1 shows the OAI-ORE conceptual data model that we adapted for representing resources in HydroShare.

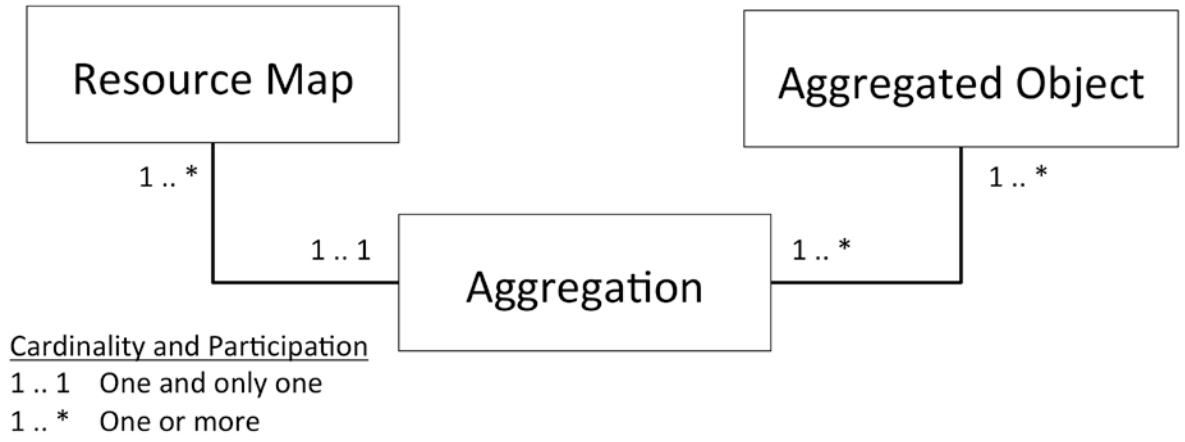

Figure 1. OAI-ORE conceptual data model for HydroShare resources. A Resource Map document describes a single Aggregation. An Aggregation is a list of one or more Aggregated Objects that are the content files of the resource. Content files can be a part of multiple Aggregations, and Aggregations can be described by more than one Resource Map.

Each resource is defined by a single resource map document, which describes an aggregation of content files. Content files are aggregated objects that may contain data, metadata, or other types of digital content. An aggregation may have one or more content files and one or more metadata files that describe the content files. One of the metadata files must be the metadata document that describes the resource as a whole. This design requires metadata at the resource-level, but does not preclude the inclusion of additional content-level metadata documents that describe one or more individual files within the aggregation. It has several advantages: 1) it provides a standard, machine readable way to describe individual content files and aggregations of content files, which makes it easier for HydroShare to automatically manage and manipulate resources; 2) because it does not limit the data types, file types, or file formats of content files that it can represent, it can be used by HydroShare to represent diverse resource types; 3) resources can also be aggregated into collections that then become

new resources; and 4) specific relationships can be maintained between content files and the metadata that describe them, which means that HydroShare can maintain relatively complex structure and relationships within a file-based archive.

The heterogeneity in file types, formats, and potential hierarchical structure of content to be shared in HydroShare required a file-based data model and drove the selection of technologies used by the resource data model. The selection of OAI-ORE for HydroShare resources was also heavily influenced by the fact that other major environmental cyberinfrastructure projects are using OAI-ORE to represent data packages, including the Data Observation Network for Earth (DataONE) project (http://www.dataone.org; DataONE, 2015) and the Sustainable Environment Actionable Data (SEAD) project (http://sead-data.net; Myers et al., 2014). In its first phase, DataONE treated datasets as opaque objects (i.e., files whose format, syntax, or structure is unknown) and did not require a specific file format or syntax for submitted data (DataONE, 2015). This model is flexible in that a DataONE package can accommodate any file-based dataset. However, processing of the data with operations such as translations, extraction or subsetting, and merging with other datasets is not well supported because little is known about the structure and syntax of content files. This type of functionality is planned for subsequent phases of the DataONE project (DataONE, 2015) and would be much easier to implement if data packages conformed to well-specified content data models that define the structure, syntax, and semantics of the datasets contained within them.

HydroShare adopted DataONE's flexible representation of "data packages" (equivalent to "resources" in HydroShare), but extended it to remove the assumption of

opacity for most resource types. This has two major advantages. First, HydroShare's adoption of DataONE's definition of data packages for HydroShare resources ensures that resources published in HydroShare are compatible with DataONE. This facilitates a goal of the HydroShare system to broaden the impact of published data and model resources by exposing them as data packages to the DataONE system. Second, by adopting specific, standardized, structured formats for each resource type, new tools for creating, visualizing, analyzing, transforming, subsetting, and integrating resources can be built both within the HydroShare website and as external software tools that interact with the HydroShare system. The availability of software tools for creating and collaborating around resources is one of the incentives for scientists to use HydroShare. We envision that software tools can promote best practices that elevate the quality and reproducibility of hydrologic research.

Resource-Level Metadata

Each resource in HydroShare is described by a resource-level metadata document that details the properties of the resource as a whole. These metadata are created by the user and are used to enable discovery and to facilitate interpretation of the resource by other users. The resource-level metadata are also used to populate the view of a landing page for each individual resource in HydroShare's website. The Dublin Core Metadata Initiative has addressed the need for a high-level description of electronic resources by providing a simple, 15-element standard metadata element set (DCMI, 2012; Weibel, 1997). At a minimum, the resource-level metadata for every HydroShare resource, regardless of its type, contains the metadata elements defined by Version 1.1

of the Dublin Core metadata element set. For each HydroShare resource type, a resource content data model may define additional required and/or optional metadata specific to a resource type that go beyond the standard Dublin Core element set (see Section 3.3 and examples in Section 6), effectively creating a Dublin Core Application Profile (Coyle and Baker, 2009) for each resource type.

HydroShare encodes the resource-level metadata for storage on disk and for transfer over the Internet (e.g., when a resource is downloaded) using the RDF/XML serialization format of the Resource Description Framework (RDF) (Brickley and Guha, 2014). Figure 2 shows a graphical representation of the resource-level metadata for a hydrologic time series resource. It highlights the standard Dublin Core metadata elements (Panel a), including the structural choices we have made in how to express the value of each element. Panel b of Figure 2 shows the additional metadata elements specific to the time series resource type. HydroShare resources are cataloged for discovery using the metadata that is expressed in their resource-level metadata documents. Given that the structure of each defined resource type is known by HydroShare, their content files are automatically parsed when they are uploaded to extract information for inclusion in the resource-level metadata document, which eases the burden on users for creation of resource-level metadata for known resource types.

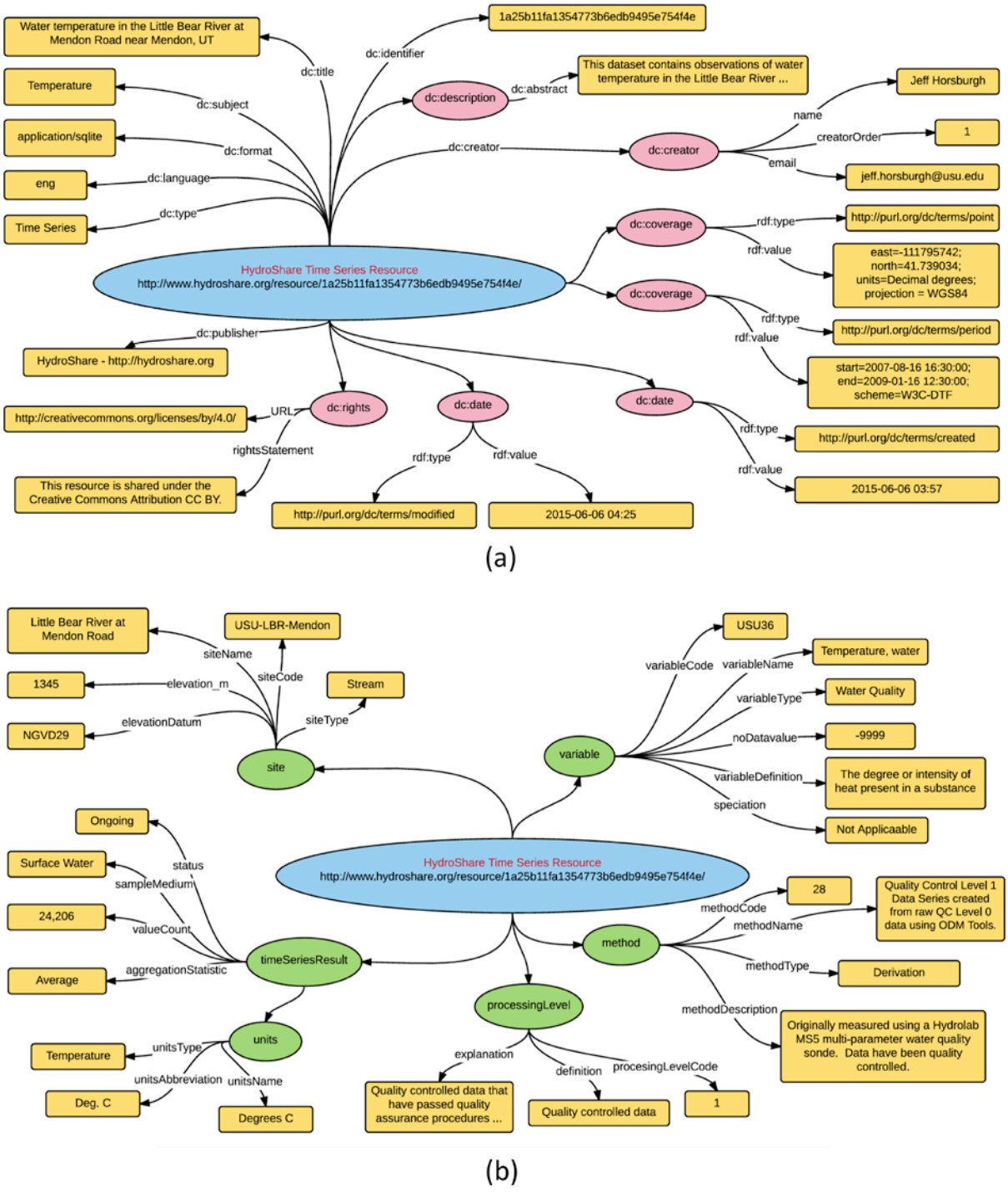

Figure 2. Graphical example of the content of resource-level metadata for a hydrologic time series resource. Panel (a) shows the standard Dublin Core metadata elements and their values. Each Dublin Core metadata element is prefixed with “dc”, and, where multiple levels of hierarchy are present, a pink node shows the first level. Panel (b) shows the extended metadata elements and values. Individual metadata element names are labeled on the arrows and their values are shown as yellow boxes.

Resource Maps

Resource map documents encode the content and structure (e.g., file hierarchies and relationships) of HydroShare resources in machine readable format that conforms to the OAI-ORE specification. The machine readability of resource map documents means that HydroShare can understand the structure of a resource by parsing its resource map, and it can validate resource contents according to rules that have been set for each resource type. OAI-ORE resource maps can be represented in one of several different machine-readable serializations such as RDF/XML, turtle, and Atom XML; detailed information about OAI-ORE resource maps can be found in the ORE User Guide (Lagoze et al., 2008b). The resource map expresses which aggregation (e.g., a collection of content files) it describes and lists the objects (the content files in the collection) that are part of the aggregation. Each object described in a resource map is identified using a web universal resource identifier (URI). HydroShare uses resolvable universal resource locators (URLs) as the URIs in resource map documents, which means that a resource's content could be recreated from its resource map in another repository or location. Because the URLs of each individual content file are resolvable, it also enables the creation of resources that reference files in other resources (e.g., rather than making copies).

Although resource maps are capable of encoding semantic relationships among aggregated objects, they do not prescribe a specific data model for the objects they describe. Relationships among the aggregated objects are provided using RDF predicates, which means that multiple types of relationships can be defined. OAI-ORE provides specific predicates that define the relationship types between the aggregation

and a resource map ("ore:describes" and conversely "ore:isDescribedBy") and between the aggregation and the objects that it aggregates ("ore:aggregates" and conversely "ore:isAggregatedBy"). However, OAI-ORE does not specifically describe how relationships between objects in the aggregation might be expressed (e.g., the relationship between the resource-level metadata file and the aggregation of content files that it describes). The DataONE project has suggested a solution (DataONE, 2012) that extends OAI-ORE to specifically include expression of the relationships between aggregated objects using the Citation Typing Ontology (CiTO, http://purl.org/spar/cito/). An example visualization of the relationships and attributes expressed in a resource map document is shown in Figure 3 for a hydrologic time series resource.

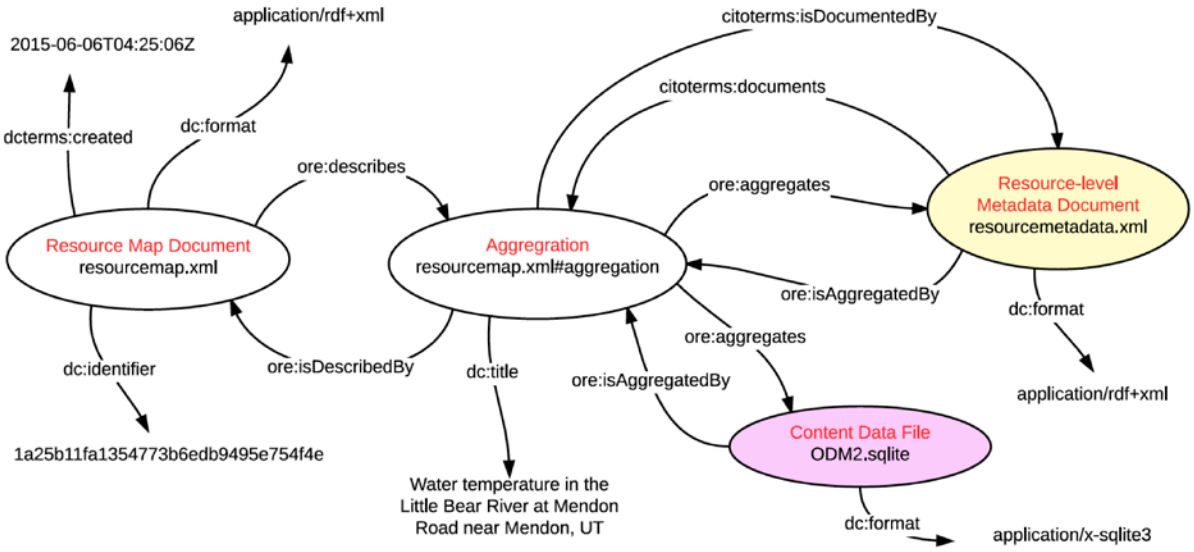

Figure 3. Example OAI-ORE representation of the structure of a HydroShare hydrologic time series resource. Time series resources have a single content data file (in this case a file named ODM2.sqlite) and a single resource-level metadata file (resourcemetadata.xml). Both of these files participate in the Aggregation, which is described by the Resource Map document (resourcemap.xml).

Because the OAI-ORE data model is flexible and general purpose, we placed additional constraints on its implementation to meet the needs of HydroShare and to ensure compatibility of HydroShare resources with DataONE. First, HydroShare uses the RDF/XML serialization for resource map documents (Lagoze et al., 2008c). Second, the URI of the aggregation object is expressed as a hash URI based on the URI of the resource map document, as recommended by ORE. This ensures that the aggregation can be referenced directly in other resource maps and still be resolved (i.e., in the case that a single set of aggregated files is described by more than one resource map document). Next, when referencing a HydroShare resource in a separate resource map document (e.g., in the case of creating a composite resource that aggregates more than one existing HydroShare resource), the URI of the resource being referenced must resolve to a resource map document. Although the HydroShare website does not yet enable users to create composite resources, this functionality is planned for future releases to enable creation of collections, larger datasets, or complex resources that may contain several other resources of different types.

Each resource in HydroShare is described with a “dc:identifier” field from the Dublin Core XML namespace in its resource map document containing the globally unique and persistent HydroShare identifier for the resource. When expressing identifiers in a URI, they are URL encoded (i.e., http://www.hydroshare.org/resource/XXXXXX where “XXXXXX” is the identifier). When expressing identifiers in the “dc:identifier” field, they are not (i.e., they are encoded simply as the text of the identifier – “XXXXXX”).

Finally, the relationship between the resource-level metadata document and the aggregation is indicated using terms drawn from the CiTO ontology. Specifically,

“citoterms:documents” is attached to an aggregated resource-level metadata file with the object of the triple being the URI of the aggregation and the converse indicated by “citoterms:isDocumentedBy.” These terms can also be used to document and enable HydroShare to automatically discover relationships among aggregated files (e.g., in the case where a content-level metadata document is included that describes one or more of the individual content files). Where needed, HydroShare has adopted additional semantic terms required to express the types of relationships among aggregated objects within a resource or relationships between two resources. For example, HydroShare needs terms to describe formal relationships among a Model Instance resource and the associated Model Program resource that is able to execute the Model Instance (“ExecutedBy”). Encoding these relationships is important to ensure that the structure and relationships among objects within a resource and those between resources can be automatically interpreted by a machine.

Content Data Models for HydroShare Resources

HydroShare does not prevent users from creating resources containing objects that are of types unknown to the system (similar to the DataONE model), but treats these as “generic” resources. No value added functionality is provided for generic resources other than allowing users to create them, describe them with standard metadata, set access control permissions on them, version them, share them, comment on them, rate them, and download them. In contrast, we are now developing value-added tools for a standard set of well-known hydrologic resource types. The list of standard resource types currently supported by HydroShare is listed in Table 1; however, we anticipate

adding several other resource types, including geographic features, river geometry, and sample-based observations.

Table 1. Resource types currently supported by HydroShare.

Resource Type	Description
Generic	A package of one or more files for which HydroShare does not know the specific type.
Time Series	Time series of hydrologic observations from point monitoring sites, including streamflow gages, water quality monitoring stations, weather stations, etc.
Referenced Time Series	A link to a URL endpoint that represents a time series dataset hosted on an external CUAHSI HIS HydroServer.
Geographic Raster	Georeferenced grids containing datasets such as land cover, elevation, elevation derivatives, etc.
Multidimensional Space/Time Dataset	Continuous space/time grids such as radar-based rainfall data.
Model Program	The computational engine for a model. Consists of source code or compiled software for executing the model and software related metadata such as version, language, platform, etc. Model programs are not place-based.
Model Instance	A set of files describing a simulation model constructed for a specific place and time. Model instances can have specific metadata, input files, and output files.

Resource content data models have been developed for each of the standard resource types listed in Table 1 that specify: 1) data content, structure, and format; 2) the name and type of all data and metadata elements; 3) which metadata elements are required or optional; and 4) file formats for import, storage, and export. It is beyond the scope of this paper to describe each of these resource content data models in detail; however, Table 2 provides an example summary specification for the hydrologic time series resource type. Resource content data models may also specify the use of controlled vocabularies for the content of standard Dublin Core or extended, resource-specific metadata elements. For example the time series resource uses Version 2 of the

Table 2. Example summary resource content data model specification for the hydrologic time series resource type.

Content Data Model Element	**Description**
Name	Time Series Resource
Data Content and Structure	Time series resources conform to Version 2 of the Observations Data Model (http://www.odm2.org).
Metadata Elements Beyond Dublin Core	Site Information: SiteCode, SiteName, Elevation, Elevation Datum, SiteType Variable Information: VariableCode, VariableName, VariableDefinition, VariableType, NoDataValue, Speciation Method Information: MethodCode, MethodName, MethodType, MethodDescription, MethodLink Processing Level: ProcessingLevelCode, Definition, Explanation Result Information: UnitsType, UnitsName, UnitsAbbreviation, Status, SampledMedium, ValueCount, AggregationStatistic, ValueCount
Internal Storage Format	Relational database in SQLite (https://www.sqlite.org).
Import Formats	Comma Separated Values (CSV) file, WaterML file, YAML Observations Data Archive (YODA) file
Export Formats	SQLite database, CSV file, WaterML file, YODA file

Observations Data Model (http://www.odm2.org) for which extensive controlled vocabularies have been developed (http://vocabualry.odm2.org). In an effort to keep the barrier for creating content in HydroShare low for users, HydroShare does not currently validate resource metadata created by users against controlled vocabularies and reject resources that are not compliant. However, future functionality may be added to assist users during the metadata creation process to encourage the use terms from existing controlled vocabularies where they are available.

Adding a new, standard resource type to HydroShare requires development of resource-type specific tools that enable users to open, visualize, convert, analyze, and otherwise manipulate the contents of resources beyond standard create, read, update,

delete, and social interaction functionality of generic resources. HydroShare encourages users to adopt the content data models for resource types that are supported and for which functionality has already been developed. The motivation for users is that HydroShare provides value added tools (e.g., visualization, processing, analysis, transformation) for supported resource types, whereas HydroShare treats unknown resource types as opaque objects with no such functionality provided.

PACKAGING RESOURCES

HydroShare uses the BagIt File Packaging Format (Boyko et al., 2012) for storing resources on disk and for serializing them to zipped files that can be transferred over the Internet (e.g., when a HydroShare user requests to download a resource). BagIt is a hierarchical file packaging format designed specifically for storage and transfer of digital content and has been used in several library and digital curation implementations (e.g., Cramer and Kott, 2010). A “bag” consists of arbitrary content (e.g., files) and “tags,” which are metadata files that document the contents of the bag. BagIt does not attempt to provide a data model for the data it carries, instead asserting that consuming software applications (in this case the HydroShare web application and HydroShare client applications) will know how to handle the contents of the bag based on the metadata included within it or via prior knowledge of its contents. Bags are ideal for digital content normally kept as a collection of files and are also well suited for export or archival purposes for content normally kept in database structures. Standardization of disk storage and network transport for HydroShare resources using BagIt enables external client functions that operate on HydroShare resources to be developed. The

HydroShare web service API and access control model allow external clients, which can be developed independently from HydroShare, to create and retrieve bags, using HydroShare as a storage resource for content creation.

Resource Bag Structure

A bag consists of a base directory that contains a set of required and optional tag files, a sub-directory named “data,” which is called the “payload directory” and within which the content files are stored, and a set of optional tag directories. The payload files in the data directory are an arbitrary file hierarchy. The tag files in the base directory consist of a file named “manifest-md5.txt”, a file named “bagit.txt,” and zero or more additional tag files. The tag files in the optional tag directories are also arbitrary file hierarchies, and the tag directories may have any name that is not reserved for a file or directory in the BagIt specification. In the BagIt specification, the base directory may have any name, but HydroShare uses the resource identifier to name the base directory. The base directory and all of the files and directories it contains are stored as a directory/file hierarchy on disk and not within a database management system. Figure 4 shows an example of this structure for the hydrologic time series resource whose resource map is shown in Figure 3 above.

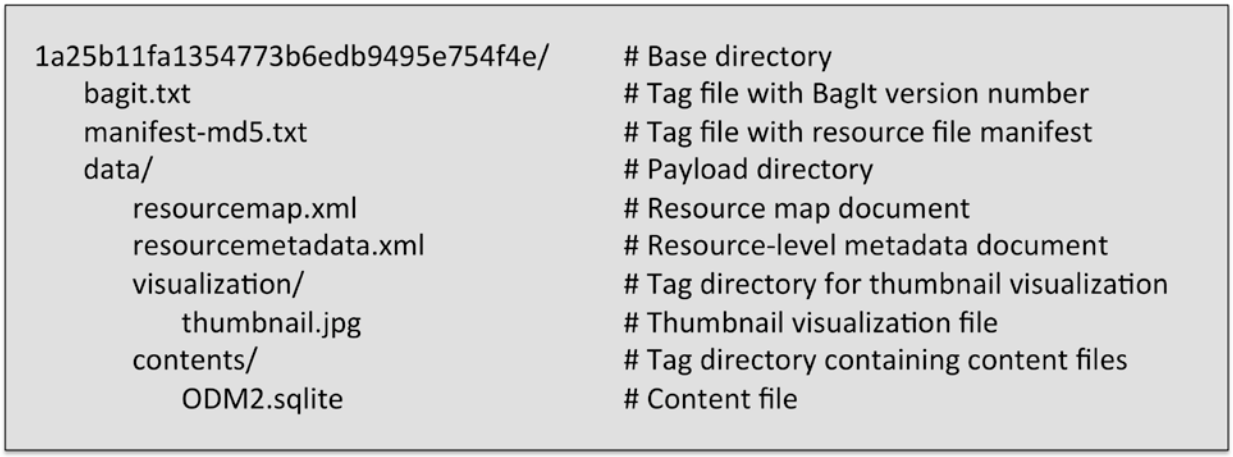

Figure 4. Example of the HydroShare implementation of the BagIt directory structure for a hydrologic time series resource. The base directory for the bag is named using the HydroShare identifier for the resource. The content of the resource is stored in a hierarchical file/directory structure within the base directory.

HydroShare uses all of the required elements of the BagIt specification. The "bagit.txt" file in the base directory contains two lines that define the BagIt version and character encoding of the tag files. This file ensures that the HydroShare system and client software can automatically detect the BagIt version (currently Version 0.96 at the time of this writing) used by the resource and handle the contents of the bag appropriately. The "manifest-md5.txt" tag file in the base directory lists payload files and checksums for those files generated using the md5 algorithm. The payload manifest asserts integrity of the payload in a bag using checksum algorithms. Each line in the payload manifest file contains the checksum and filename for an individual file within the resource, where the checksum is a hex-encoded checksum calculated over every octet in the file and the filename is the pathname of the file relative to the base directory. For HydroShare, this is important from an archival and storage management perspective. When bags are moved from one storage resource to another (e.g., in the event of a

future upgrade of the hardware on which HydroShare is hosted), the checksums in the manifest file can be used to verify the integrity of the content files after they have been moved to the new storage resource. Additionally, the integrity of the content files can be verified by client software by inspecting the checksums after downloading resources from HydroShare.

The “data” directory may contain any number of sub-directories. The files under the “data” directory are called the “payload files,” or the “payload.” The payload is treated as octet streams (e.g., binary files) for all purposes relating to the BagIt specification, and is not otherwise prescribed by the BagIt specification. However, in HydroShare the specific structure of the file/directory hierarchy is defined by the content data model for a resource type. Every resource, regardless of type, will contain a resource map document (resourcemap.xml) and a resource-level metadata document (resourcemetadata.xml) in the “data” directory. A “visualization” directory is an optional tag folder that may contain a thumbnail visualization of the resource (thumbnail.jpg). HydroShare uses these thumbnail images for a preview display of the resource in its landing page. Finally, every HydroShare resource contains a “contents” directory within the “data” directory, within which the content files are stored. In the example for a hydrologic time series resource above, a single content file (ODM2.sqlite) containing the time series data is present in the “contents” directory.

Serialization of Resource Bags for Transport

When users request to download a resource either through the HydroShare website or via the HydroShare web service application programming interface (API),

HydroShare serializes the resource bag's file system hierarchy (i.e., the base directory) into a single-file archive format ZIP file. HydroShare conforms to the BagIt specification rules for bag serialization. For serialized resources, the top-level directory contains a single bag, and the serialized ZIP file is named with the same name as the bag's base directory (e.g., when a user downloads a file named "b2e4b18dd8654ab4b508d32ef2380129.zip" it will unzip to a base directory named "b2e4b18dd8654ab4b508d32ef2380129"). Bags are serialized from their parent directory to ensure that when a serialized bag is unzipped a single base directory is created that contains all of the payload and tag files.

RESOURCE CREATION WORKFLOW

New resources are created within HydroShare via a number of different mechanisms. Currently supported methods include: 1) uploading content files and metadata through the HydroShare website; 2) specifying link(s) for web resources that are hosted elsewhere (e.g., referenced time series hosted on a remote server); and 3) uploading content files and metadata through a web service API from client software. Future functionality will include creating new resources by computational operations on resources such as subsetting or evaluation of a derivative quantity (e.g., slope from a digital elevation model). HydroShare attempts to automatically harvest as much metadata as possible from uploaded or remote sources by parsing the content files. Any required metadata elements that cannot be automatically harvested by HydroShare are left to be input or added later by a user. Regardless of the mechanism by which new resources are created within HydroShare, the HydroShare system completes the

actions as shown in Figure 5 for each new resource as part of the resource creation workflow.

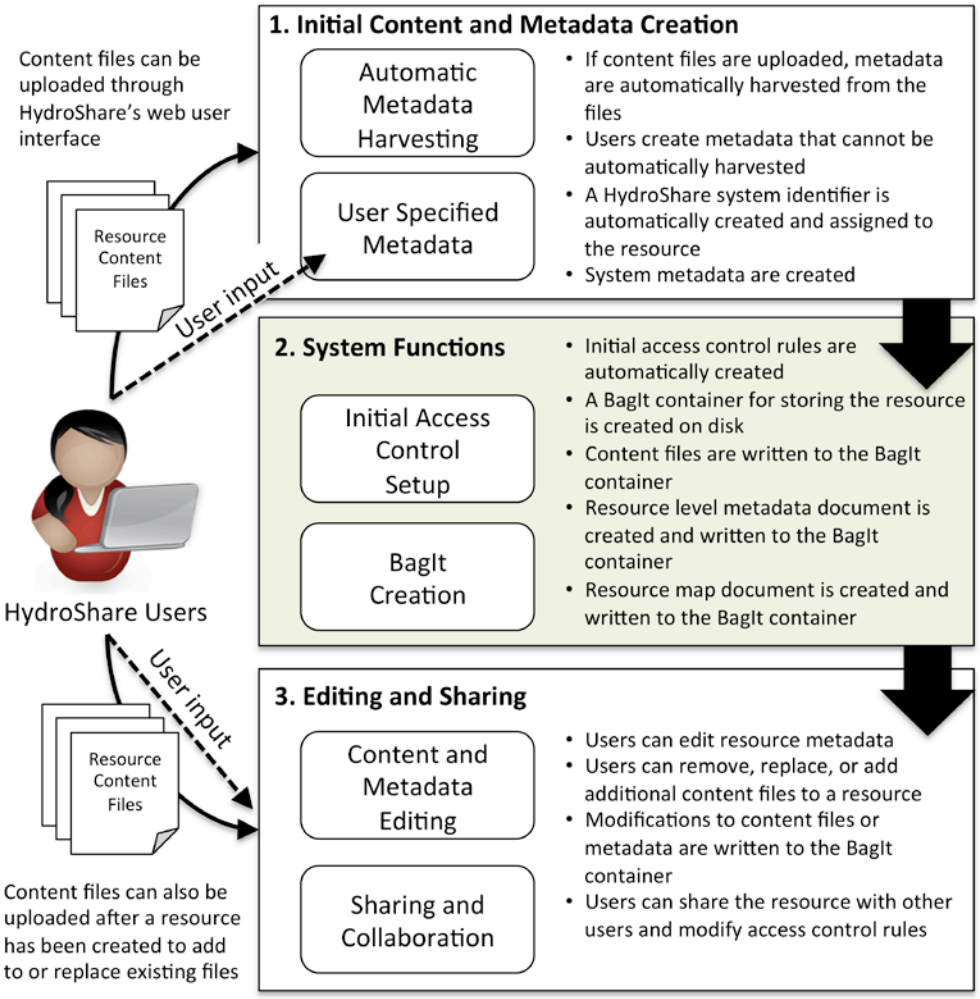

Figure 5. HydroShare resource creation workflow showing actions completed for each new resource.

EXAMPLE HYDROLOGIC RESOURCES

Single Content File Resources

Several of the currently supported HydroShare resource types are based on content data models that encapsulate the data within a single content file. These include time series, geographic rasters, and multidimensional space/time datasets. The structure of

all single content file resources is nearly identical to that shown in Figures 3 and 4 for the time series resource. However, the resource types differ in the format of the single content file and the additional metadata elements that extend their resource-level metadata documents. Details of how a hydrologic time series dataset is represented using the HydroShare resource data model are shown in the tables and figures above. The data values of the time series are stored in a single SQLite database file that conforms to the ODM2 data model. Figures 3 and 4 and Table 2 are based on a water temperature time series dataset for the Little Bear River at Mendon Road near Mendon, Utah that serves as a fully specified example hydrologic time series resource and was published in HydroShare by Horsburgh and Jones (2015). Like time series, geochemistry/sample-based datasets will also use the ODM2 data model, have very similar extended metadata elements, and will be stored in a single SQLite database.

A geographic raster resource consists of a georeferenced grid for representing imagery, digital elevation models, or other grid-based data products common in hydrology such as land cover, elevation derivatives like slope and aspect, etc. We decided to standardize on the GeoTIFF file format in HydroShare for raster file uploading and internal data storage because it is a public domain standard. A geographic raster resource currently consists of a file in GeoTIFF format, which contains the data with single or multiple bands and associated georeferenced metadata tags in a single content file. Additional metadata elements extracted from the GeoTIFF file for inclusion in the resource-level metadata document include “Spatial Reference,” “Cell Information,” and “Band Information.” The “Spatial Reference” element includes as sub-elements the spatial extent and coordinate system for georeferencing purposes.

These are expressed using the Dublin Core Box encoding scheme (Cox et al., 2006). The “Cell Information” element includes information describing the raster cells, including the number of rows and columns and cell size. The “Band Information” element includes information about each raster layer, or band included in the dataset. Tarboton (2015a, 2015b) has published a digital elevation model and used it to derive specific catchment area for the Logan River Watershed. These serve as fully specified examples of geographic raster resources, including relationships “DerivedFrom” and “IsDataFor” that link them.

A multidimensional (NetCDF) resource consists of a dataset stored in the Network Common Data Form (NetCDF) format to represent geographic gridded data that may have temporal or other dimensions such as altitude, pressure levels, etc. Like the GeoTIFF file used for the geographic raster resource, the NetCDF format contains the multidimensional data and descriptive metadata in a single content file. Additional metadata elements extracted from the NetCDF file for inclusion in the HydroShare resource-level metadata document include “Spatial Reference” and “Variable” information. The “Spatial Reference” element includes a definition of the spatial extent and coordinate system for the dataset. The “Variable” element describes the data variables in the file for the purpose of data reuse. As a fully specified example of the multidimensional space and time resource type, Gichamo (2015) published a multidimensional (NetCDF) resource in HydroShare that is an output from the Utah Energy Balance snowmelt model (Tarboton and Luce, 1996). This resource contains snow water equivalent data that varies in both space (i.e., latitude and longitude) and time dimensions.

Multiple Content File Resources

In addition to single content file resources, there is also a need to represent data and metadata for resources that are composed of multiple content files. This includes resources whose content cannot easily be aggregated into a single file (e.g., a geospatial dataset stored in the multi-file, ESRI shapefile format) or those that consist of both required and optional files. In contrast to single file resources, multi-file resources may have content that requires metadata in addition to or more specific than the resource-level metadata. There may also be internal relationships among content files within a resource that must be expressed or relationships with other resources in HydroShare. Thus, multi-file resources must consider not only how to associate metadata with individual files, but also how to maintain internal relationships that may exist between individual files. One example of this concept is the representation of computational hydrologic models.

As stated earlier, we conceptualize computational models as two separate but related resources within HydroShare: Model Programs and Model Instances. Model Programs describe the computational engine for the model and are not place-based. They consist of the source code or compiled software for executing the model as well as specific metadata to describe the software. Model Instances comprise a set of files describing a simulation model constructed for a specific place and time, and can have specific metadata, input files, and output files. These two resources are related in that a Model Instance is executed by a Model Program. However, they are considered separate resources in HydroShare because a single Model Program resource can be used to execute many different Model Instance resources, which are implementations of

the model for different places and times (Morsy et al., 2014). These design considerations satisfy resource-level relationships and cardinality. For example, one Model Program resource may be related to many Model Instance resources. Each of these Model Instance resources consists of unique metadata, files, and relationships that describe the hydrologic model and its simulation.

The content data model for a Model Instance resource was designed to describe a hydrologic model run. Therefore, it must consider both required and optional input and output files and was designed to remain general so that it could be used to store a run generated by any Model Program. As an example, consider a Model Instance resource for storing one of the model runs presented in Morsy et al. (2013) for studying flooding within an urbanized watershed in Columbia, South Carolina. This Model Instance resource would be associated in HydroShare with the Storm Water Management Model (SWMM) Model Program resource using the “ExectuedBy” relationship.

To share this Model Instance in HydroShare, the files associated with the Model Instance were aggregated using OAI-ORE specifications as illustrated in Figure 6. In this case, the Model Instance consists of multiple input files (i.e., *.ini, *.inp, *.hsf, etc.) as well as multiple output files (i.e., *.rpt, *.out, etc.), all of which are stored on disk and referenced by the resource map and BagIt files (Figure 7) that are automatically generated.

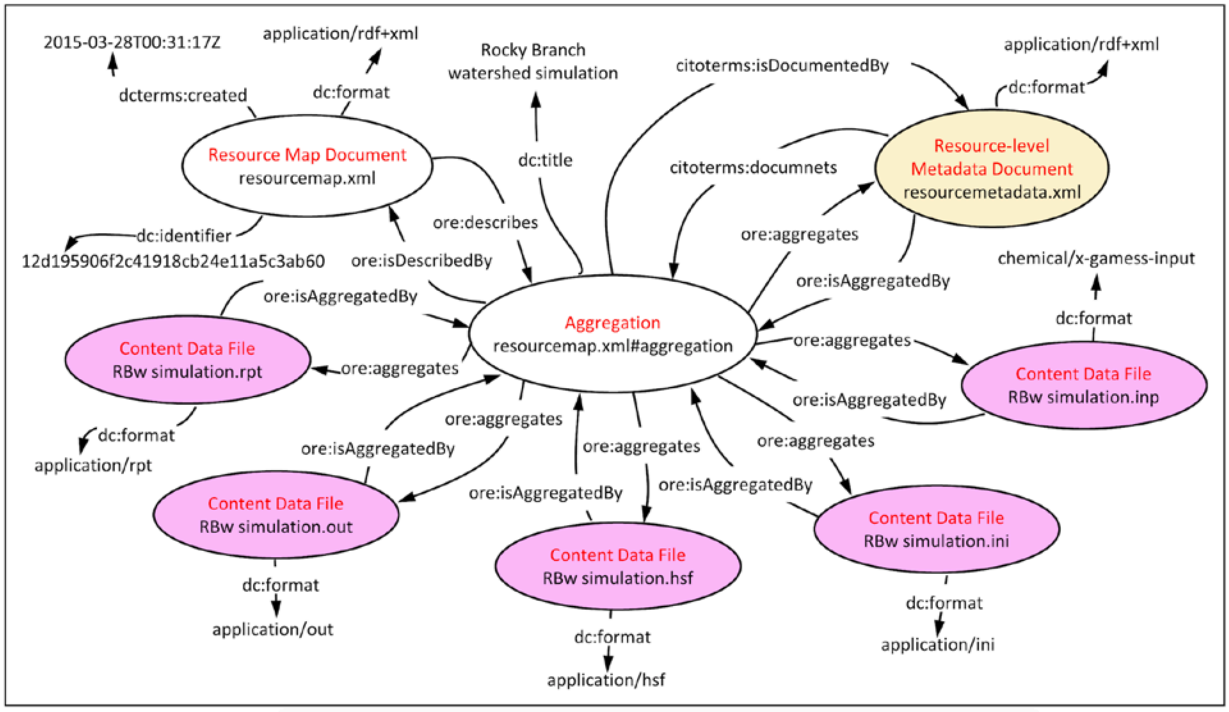

Figure 6. Graphical representation of the OAI-ORE structure for a HydroShare Model Instance resource. This resource consists of five content data files and the resource-level metadata document that participate in the Aggregation.

```
12d195906f2c41918cb24e11a5c3ab60/          # Base directory
    bagit.txt                              # Tag file with BagIt version number
    manifest-md5.txt                       # Tag file with resource file manifest
    data/                                  # Payload directory
        resourcemap.xml                    # Resource map document
        resourcemetadata.xml               # Resource-level metadata document
        visualization/                     # Tag directory for thumbnail visualization
        contents/                          # Tag directory containing content files
            RBw simulation.hsf             # Content file (hot start file)
            RBw simulation.ini             # Content file (settings file)
            RBw simulation.inp             # Content file (input file)
            RBw simulation.out             # Content file (output file)
            RBw simulation.rpt             # Content file (report file)
```

Figure 7. An example of the BagIt directory structure for a HydroShare Model Instance resource.

The files within the Model Instance resource have internal relationships that can be described within a multi-file resource (Figure 8). For example, an input file (*.inp) “uses” the settings (*.ini) and hotstart (*.hsf) files. An input file can also be used to “generate” output (*.out) and report (*.rpt) files. Lastly, a report (*.rpt) file is “derived from” an output (*.out) file. These relationships must be represented as generically as possible to remain applicable to a wide range of hydrologic models, while still conveying how the contents of a multi-file resource are interrelated. This enables multi-file relationships to be leveraged by software systems to visually display dependencies between files within resources. Future implementations of HydroShare could allow users to tag individual files within a multi-file resource, like the Model Instance resource type, recording such relationships as RDF triples in the OAI-ORE resource map for the resource. The SWMM Model Program (Rossman et al., 2015) has been published as a HydroShare resource, and Morsy (2015) published the Model Instance described in this example as a resource.

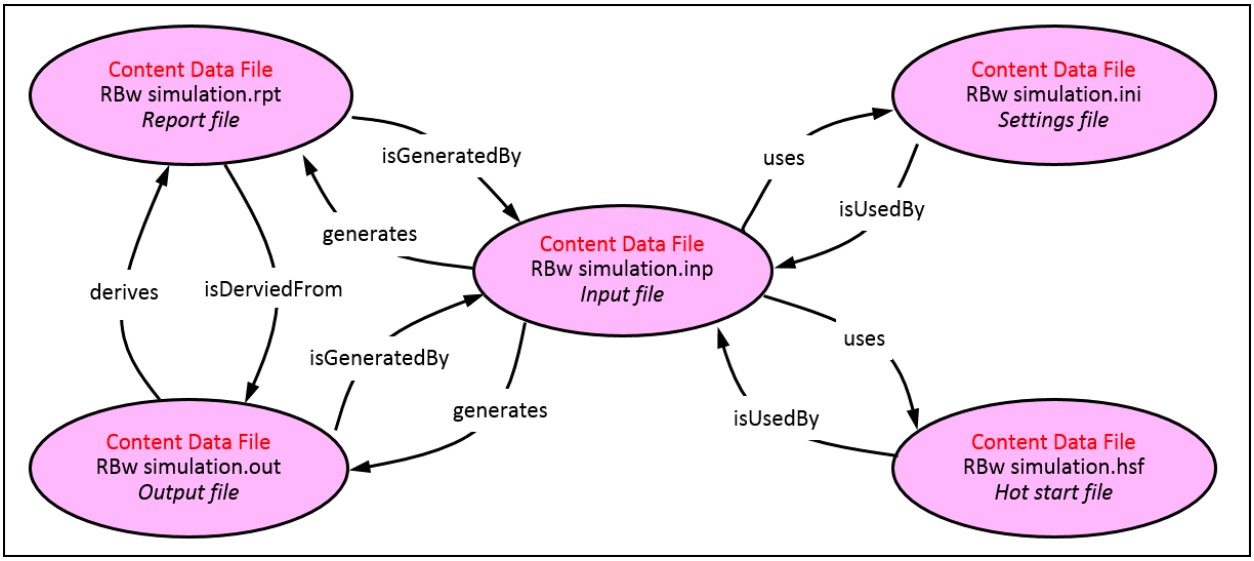

Figure 8. The relationships between the model instance resource BagIt content files. Specific semantic relationships among the files are indicated (e.g., "generates," "isGeneratedBy"). These relationships are formalized and encoded in a machine interpretable syntax within the Resource Map document for the resource.

SUMMARY AND FUTURE WORK

We developed a standardized resource data model and packaging scheme for hydrologic data and model resources by adopting and adapting existing technologies, including arbitrary file hierarchies, OAI-ORE, and BagIt. HydroShare uses this data model to store and manage dataset and model resources common within the hydrologic science community, as well as for transporting them over the Internet. It also ensures that HydroShare can consistently catalog resource metadata for supporting data discovery. The flexibility of HydroShare's data model does not constrain the types of resources that it can describe; however, where value added functionality is needed for a particular resource type, a resource content data model is required to specify the structure, syntax, and content of different resource types. We have already developed

resource content data models for many of the most common data types used in the hydrology community as well as specific functionality for managing them within HydroShare.

HydroShare's resource data model and packaging scheme are not specific to hydrologic data types. The use of OAI-ORE for capturing the structure of complex, file-based datasets and BagIt for packaging them for storage on disk and transport over the Internet would support datasets from many scientific domains. Although we have described how we have applied these technologies for hydrologic data types, the approach we used leaves flexibility for defining specific structure and semantics for the content of domain-specific resources by allowing the definition of an appropriate resource content data model.

The resource data model described here explicitly captures the information needed for both computers and users to interpret the content of a resource. The resource-level metadata document describes the resource using the standard Dublin Core metadata elements, with extensions for specific resource types, and is encoded in machine readable format. The resource map document lists the content files within the resource and semantically expresses any relationships among them, providing a computer with a way to discover a resource's structure. System metadata for a resource include time stamps for creation and modification and access control rules, and social metadata include ratings, annotations, and comments made that increase the knowledge content of resources. The storage and zip file serialization of a resource using BagIt are flexible for use within any arbitrary file hierarchy but ensure the integrity of resources using checksums recorded in a resource's manifest file.

The advantages of this approach for a collaboration system like HydroShare are that storage on disk, access control, serialization for transport over the Internet, and cataloging for discovery purposes can all be done consistently at the resource level, regardless of the resource type or content. This simplifies many aspects of the HydroShare website and back-end data store because all resources can be handled in the same way. Furthermore, the additional specification of resource content data models for known resource types enables the creation of value-added visualization and analysis tools that operate on specific resource types. The HydroShare resource data model is also consistent with the way DataONE packages datasets, which means that published HydroShare resources could be exposed for cataloging by DataONE. We believe that this will be attractive for many HydroShare users within the hydrologic science community who want to broaden the impact of their published datasets by exposing them broadly to systems like DataONE, but who need the tools offered by HydroShare for creating and describing publishable content.

Future work with HydroShare will include adding additional social functionality around resources, including the ability to build collaborative groups and to create and share resources within these social groups. The first set of social functions (e.g., sharing, access control, ratings, and comments) was implemented for HydroShare resources using design patterns consistent with those that we saw on other social media websites. To our knowledge, a single standard for social metadata has not emerged, particularly with respect to data and models as social objects. We anticipate that the work we have already done will help set the stage for standards to emerge. Much of the social metadata HydroShare collects about resources will consist of free

text comments and annotations written by HydroShare users for human interpretation. However, new opportunities and challenges exist for automating the extraction of machine interpretable information about resources from this social metadata. Topic modeling (e.g., Taurob et al., 2012; Tuarob et al., 2015) is one approach that shows promise for automatically extracting keywords, subject areas, or topics for which a resource is relevant from free text user annotations to improve data discovery. We plan to explore these approaches as we build a larger corpus of resources within HydroShare.

Given that the number and types of resources within HydroShare are growing, we are currently investigating approaches for implementing advanced data discovery functionality beyond the simple filtering of resources currently available in HydroShare. Data discovery is dependent upon the metadata stored within each resource, and this was a major driver for HydroShare to adopt a single standard for the base metadata elements for all resources (i.e., so they could be cataloged consistently). We anticipate supporting discovery queries by space, time, and keywords, similar to the discovery functions of the CUAHSI HIS; however, we are also investigating opportunities for more advanced data discovery and filtering based on resource type, similarity to other resources, and potentially additional attributes extracted from resources' extended or social metadata.

Future work will focus on merging of the model metadata framework built for HydroShare along with other existing model metadata frameworks. In particular, Elag and Goodall (2013) proposed a Water Resources Component (WRC) ontology framework specific for describing model components, and Peckham (2014) proposed an

approach for creating standard names for variables and assumptions used in geoscience modeling. While there are distinct perspectives across these past studies, it is likely that there are also opportunities for synthesizing and harmonizing the model metadata framework presented in this work with these other frameworks. Moving toward a single ontology for water resources modeling will be important step for supporting interoperability across systems.

We are also working to interface HydroShare with computational resources for executing Model Instances. This will enable scientists to execute models shared by their colleagues and reproduce results without needing to first create an appropriate simulation environment on their own local computer. SWATShare (Rajib et al., 2014), which is a web portal for publishing, sharing and running models developed using the Soil and Water Assessment Tool (SWAT), is one example of such an environment. A resource bag provides a convenient structure for transporting Model Instance resources to a computational resource like SWATShare and then for packaging results for sharing as a new resource in HydroShare. Transport of resources between systems will be accomplished via HydroShare's web service APIs, which currently have basic functionality for creating and accessing resources but are under development to add more advanced functionality. Finally, we are exploring partnerships with existing DataONE Member Nodes for exposing published HydroShare resources to the DataONE network in efforts to interoperate with other major data networks.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation under collaborative grants ACI 1148453 and 1148090 for the development of HydroShare (http://www.hydroshare.org). 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.

LITERATURE CITED

Boyko, A., J. Kunze, J. Littman, L. Madden, B. Vargas (2012). The BagIt File Packaging Format (v0.97), Network Working Group Internet Draft, available at http://tools.ietf.org/html/draft-kunze-bagit-10, last accessed 2/20/2015.

Brickley, D., R.V. Guha (2014). RDF Schema 1.1, W3C Recommendation, http://www.w3.org/TR/2014/REC-rdf-schema-20140225/, last accessed 5/9/2015.

Costello, M.J., W.K. Michener, M. Gahegan, Z.Q. Zhang, P.E. Bourne (2013). Biodiversity data should be published, cited, and peer reviewed, Trends in Ecology & Evolution, 28(8), http://dx.doi.org/10.1016/j.tree.2013.05.002.

Costello, M.J., J. Wieczorek (2014). Best practice for biodiversity data management and publication, Biological Conservation, 173, 68-73, http://dx.doi.org/10.1016/j.biocon.2013.10.018.

Coyle, K., T. Baker (2008). Guidelines for Dublin Core Application Profiles, Dublin Core Metadata Initiative, http://dublincore.org/documents/profile-guidelines/, last accessed 6/2/2015.

Cox, S., A. Powel, A. Wilson, P. Johnson (2006). DCMI Box Encoding Scheme: Specification of the spatial limits of a place, and methods for encoding this in a text string, Dublin Core Metadata Initiative, http://dublincore.org/documents/dcmi-box/, last accessed 5/9/2015.

Cramer, T., K. Kott (2010). Designing and implementing second generation digital preservation services: A scalable model for the Stanford Digital Repository, D-Lib Magazine, 16 (9/10), http://dx.doi.org/10.1045/september2010-cramer.

DataONE (2015). DataONE Architecture, Version 1.2, available at http://releases.dataone.org/online/api-documentation-v1.2.0/, last accessed 3/7/2015.

De La Beaujardiere, J. (2008). The NOAA IOOS Data Integration Framework: Initial implementation report, OCEANS 2008, 15-18 Sept. 2008, p. 1-8, http://dx.doi.org/10.1109/OCEANS.2008.5152007.

Dublin Core Metadata Initiative (DCMI) (2012). DCMI Metadata Terms, http://dublincore.org/documents/dcmi-terms/, last accessed 5/9/2015.

Dublin Core Metadata Initiative (DCMI) (2009). Guidelines for Dublin Core Application Profiles, http://dublincore.org/documents/profile-guidelines/, last accessed 5/9/2015.

Dunlap, R., L. Mark, S. Rugaber, V. Balaji, J. Chastang, L. Cinquini, C. DeLuca, D. Middleton, S. Murphy (2008). Earth system curator: metadata infrastructure for climate modeling, *Earth Science Informatics*, *1*(3-4), 131-149, http://dx.doi.org/10.1007/s12145-008-0016-1.

Elag M., J.L. Goodall (2013). An ontology for component-based models of water resource systems, *Water Resources Research*, 29(8), 5077-5091, http://dx.doi.org/10.1002/wrcr.20401.

Engeström, J. (2005). Why some social network services work and others don't – Or: the case for object-centered sociality, http://www.zengestrom.com/blog/2005/04/why-some-social-network-services-work-and-others-dont-or-the-case-for-object-centered-sociality.html, last accessed 5/9/2015.

Gichamo, T.Z. (2015). Snow water equivalent estimation at TWDEF site from Oct 2009 to June 2010, HydroShare, http://www.hydroshare.org/resource/50fa9c5b3b5b4263bf4f1752a8b6723c.

Gregersen, J. P. Gijsbers, and S. Westen. (2007). OpenMI: Open Modeling Interface, Journal of Hydroinformatics, 9(3), 175-191, http://dx.doi.org/0.2166/hydro.2007.023.

Heard, J., D. Tarboton, R. Idaszak, J. Horsburgh, D. Ames, A. Bedig, A. Castronova, A. Couch, P. Dash, C. Frisby, T. Gan, J. Goodall, S. Jackson, S. Livingston, D. Maidment, N. Martin, B. Miles, S. Mills, J. Sadler, D. Valentine, L. Zhao (2014), An architectural overview of HydroShare, a next-generation hydrologic information system, in: Proceedings of the 11th International Conference on Hydroinformatics, HIC 2014, New York City, USA, http://www.hic2014.org/proceedings/handle/123456789/1536.

Heidorn, P.B. (2008). Shedding light on the dark data in the long tail of science, Library Trends, 57(2), 280-299, http://dx.doi.org/10.1353/lib.0.0036.

Horsburgh, J., A. Jones (2015). Water Temperature in the Little Bear River at Mendon Road near Mendon, UT, HydroShare, http://www.hydroshare.org/resource/1a25b11fa1354773b6edb9495e754f4e.

Hsu, L., R.L. Martin, B. McElroy, K. Litwin-Miller, W. Kim (2015). Data management, sharing, and reuse in experimental geomorphology: Challenges, strategies, and scientific opportunities, Geomorphology, http://dx.doi.org/10.1016/j.geomorph.2015.03.039.

Lagoze, C., H. Van de Sompel, P. Johnston, M. Nelson, R. Sanderson, S. Warner (2008a). Open Archives Initiative Object Reuse and Exchange: ORE Specification – Abstract Data Model, available at http://www.openarchives.org/ore/1.0/datamodel.html, last accessed 3/3/2015.

Lagoze, C., H. Van de Sompel, P. Johnston, M. Nelson, R. Sanderson, S. Warner (2008b). Open Archives Initiative Object Reuse and Exchange: ORE User Guide – Primer, available at http://www.openarchives.org/ore/1.0/primer, last accessed 3/3/2015.

Lagoze, C., H. Van de Sompel, P. Johnston, M. Nelson, R. Sanderson, S. Warner (2008c). Open Archives Initiative Object Reuse and Exchange: ORE User Guide – Resource Map Implementation in RDF/XML, available at http://www.openarchives.org/ore/1.0/rdfxml.html, last accessed 3/12/2015.

Lehnert, K.A., S.M. Carbotte, W.B.F. Ryan, V. Ferrini, K. Block, R.A. Arko, C. Chan (2011). IEDA: Integrated Earth Data Applications to support access, attribution, analysis, and preservation of observational data from the ocean, earth, and polar sciences, Geophysical Research Abstracts, 13, EGU2011-13113.

Michener, W., D. Vieglais, T. Vision, J. Kunze, P. Cruse, G. Janée (2011). DataONE: Data Observation Network for Earth – Preserving data and enabling innovation in the biological and environmental sciences, D-Lib Magazine, 17(1/2), http://dx.doi.org/10.1045/january2011-michener.

Morsy, M.M. (2015). Rocky Branch watershed simulation, HydroShare, http://www.hydroshare.org/resource/12d195906f2c41918cb24e11a5c3ab60.

Morsy, M.M., J.L. Goodall, C. Bandaragoda, A.M. Castronova, J. Greenberg (2014). Metadata for describing water models, in: D.P. Ames, N.W.T. Quinn and A.E. Rizzoli (eds), Proceedings of the 7th International Congress on Environmental Modelling and Software, San Diego, California, USA, International Environmental Modelling and Software Society (iEMSs), ISBN: 978-88-9035-744-2, http://www.iemss.org/sites/iemss2014/papers/iemss2014_submission_259.pdf.

Morsy, M.M., J.L. Goodall, M.E. Meadows (2013). Flood mitigation in an urbanized watershed using a combination of BMP and LID techniques, in: Proceedings of the World Environmental & Water Resources Congress 2013, ASCE, Cincinnati, Ohio, USA, May 19-23, http://dx.doi.org/10.13140/2.1.3043.1682.

Myers, J. and the SEAD Team (2014). Publishing heterogeneous and multi-source research data with rich metadata and round trip provenance: Opportunities and challenges as data services form an ecosystem, Abstract IN33C-3783 presented at 2014 Fall Meeting, American Geophysical Union, San Francisco, CA, December 15-19.

Nagai, M., A. Rajbhandari, M. Ono, M., R. Shibasaki (2012). Earth Observation Data Interoperability Arrangement with Ontology Registry, *Communications in Computer and Information Science*, 421, 128-136, http://dx.doi.org/10.1007/978-3-319-08732-0_10.

Peckham, S.D. (2014), The CSDMS standard names: Cross-domain naming conventions for describing process models, data sets, and their associated

variables, in: D.P. Ames, N.W.T. Quinn and A.E. Rizzoli (eds), Proceedings of the 7th International Congress on Environmental Modelling and Software, San Diego, California, USA, International Environmental Modelling and Software Society (iEMSs), ISBN: 978-88-9035-744-2, https://csdms.colorado.edu/mediawiki/images/Peckham_2014_iEMSs.pdf.

Peckham, S.D., E.W.H. Hutton, B. Norris (2013). A component-based approach to integrated modeling in the geosciences: The design of CSDMS, Computers & Geoscienes, http://dx.doi.org/10.1016/j.cageo.2012.04.002.

Rajib, M.A., V. Merwade, I.L. Kim, S. Zhe, L. Zhao, C. Song (2014). SWATShare – A web-portal for hydrology research and education using Soil Water And Assessment Tool, in Proceedings of the 11th International Conference on Hydroinformatics, New York City, NY, Aug. 17-21, http://www.hic2014.org/proceedings/bitstream/handle/123456789/1704/1833.pdf.

Rossman, L., T. Schade, D. Sullivan, R. Dickinson, C. Chan, E. Burgess (2015). Storm Water Management Model (SWMM), HydroShare, http://www.hydroshare.org/resource/2cddae40e9594c21b947fdbbe4225439.

Tenopir, C., R.J. Sandusky, S. Allard, B. Birch (2014). Research data management services in academic research libraries and perceptions of librarians, Library & Information Science Research, 36, 84-90, http://dx.doi.org/10.1016/j.lisr.2013.11.003.

Tarboton, D. (2015a). Logan Digital Elevation Model, HydroShare, http://www.hydroshare.org/resource/b7822782896143ca8712395f6814c44b.

Tarboton, D. (2015b). Logan Specific Catchment Area, HydroShare, http://www.hydroshare.org/resource/877bf9ed9e66468cadddb229838a9ced.

Tarboton, D.G., C.H. Luce (1996). Utah Energy Balance Snow Accumulation and Melt Model (UEB), Computer model technical description and users guide, Utah Water Research Laboratory and USDA Forest Service Intermountain Research Station. http://www.neng.usu.edu/cee/faculty/dtarb/snow/snowreptext.pdf.

Tarboton, D.G., J.S. Horsburgh, D.R. Maidment, T. Whiteaker, I. Zaslavsky, M. Piasecki, J. Goodall, D. Valentine, T. Whitenack (2009). Development of a community Hydrologic Information System. In: Anderssen, R. S., R. D. Braddock, and L.T.H. Newham (eds.) 18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation, Modelling and Simulation Society of Australia and New Zealand and International Association for Mathematics and Computers in Simulation, July 2009, pp. 988-994, ISBN: 978-0-9758400-7-8.

Tarboton, D.G., R. Idaszak, R., J.S. Horsburgh, J. Heard, D. Ames, J.L. Goodall, L. Band, V. Merwade, A. Couch, J. Arrigo, R. Hooper, D. Valentine, D. Maidment (2014a). HydroShare: Advancing collaboration through hydrologic data and model sharing, in: D.P. Ames, N.W.T. Quinn and A.E. Rizzoli (eds), Proceedings of the 7th

International Congress on Environmental Modelling and Software, San Diego, California, USA, International Environmental Modelling and Software Society (iEMSs), ISBN: 978-88-9035-744-2, http://www.iemss.org/sites/iemss2014/papers/iemss2014_submission_243.pdf.

Tarboton, D.G., R. Idaszak, J.S. Horsburgh, J. Heard, D. Ames, J.L. Goodall, L.E. Band, V. Merwade, A. Couch, J. Arrigo, R. Hooper, D. Valentine, D. Maidment (2014b). A resource centric approach for advancing collaboration through hydrologic data and model sharing, in: Proceedings of the 11th International Conference on Hydroinformatics, HIC 2014, New York City, USA, http://www.hic2014.org/proceedings/handle/123456789/1539.

Taurob, S., L.C. Pouchard, N. Noy, J.S. Horsburgh, G. Palanisamy (2012). ONEMercury: Towards Automatic Annotation of Environmental Science Metadata, 2nd International Workshop on Linked Science 2012—Tackling Big Data (LISC2012) - LinkedScience.org, http://linkedscience.org/wp-content/uploads/2012/05/lisc2012_submission_5.pdf.

Taurob, S., L.C. Pouchard, P. Mitra, C.L. Giles (2015). A generalized topic modeling approach for automatic document annotation, International Journal on Digital Libraries, 16(2), 111-128, http://dx.doi.org/10.1007/s00799-015-0146-2.

Weibel, S. (1997). The Dublin Core: A simple content description model for electronic resources, Bulletin of the American Society for Information Science, 24(1), 9-11, http://dx.doi.org/10.1002/bult.70.

Zaslavsky, I., T. Whitenack, M. Williams, D.G. Tarboton, K. Schreuders, A. Aufdenkampe (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, http://dx.doi.org/10.5060/D2NC5Z4X.