A RESOURCE CENTRIC APPROACH FOR ADVANCING COLLABORATION THROUGH HYDROLOGIC DATA AND MODEL SHARING

DAVID G TARBOTON (1), JEFFERY S HORSBURGH (1), RAY IDASZAK (2), JEFF HEARD (2), DAVID VALENTINE (3), ALVA COUCH (4), DAN AMES (5), JONATHAN L GOODALL (6), LARRY BAND (7), VENKATESH MERWADE (8), JENNIFER ARRIGO (9), RICHARD HOOPER (9), DAVID MAIDMENT (10)

(1): Civil and Environmental Engineering, Utah State University, Logan, UT 84322, USA
(2): RENCI, University of North Carolina at Chapel Hill, Chapel Hill, NC 27517, USA
(3): San Diego Supercomputer Center, University of California at San Diego, 10100 Hopkins Drive, La Jolla, CA 92093, USA
(4): Computer Science, Tufts University, Medford, MA 02155, USA
(5): Civil and Environmental Engineering, Brigham Young University, Provo, Utah 84602, USA
(6): Civil and Environmental Engineering, University of Virginia, Charlottesville, VA 22904, USA
(7): Geography, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
(9): School of Civil Engineering, Purdue University, West Lafayette, IN 47907, USA
(9): CUAHSI, 196 Boston Ave, Medford, MA 02155, USA
(10): Civil, Architectural and Environmental Engineering, University of Texas at Austin, Austin, Texas 78712, USA

HydroShare is an online, collaborative system being developed for open sharing of hydrologic data and models. The goal of HydroShare is to enable scientists to easily discover and access hydrologic data and models, retrieve them to their desktop or perform analyses in a distributed computing environment that may include grid, cloud or high performance computing model instances as necessary. Scientists may also publish outcomes (data, results or models) into HydroShare, using the system as a collaboration platform for sharing data, models and analyses. HydroShare is expanding the data sharing capability of the CUAHSI Hydrologic Information System by broadening the classes of data accommodated, creating new capability to share models and model components, and taking advantage of emerging social media functionality to enhance information about and collaboration around hydrologic data and models. In an information system such as HydroShare, the way that information is represented, and exposed to programming interfaces, is fundamental in enhancing the functionality that can be supported. One of the fundamental concepts in HydroShare is that of a resource. In HydroShare, a resource serves, as the basic unit of content. The Resource Data Model supports specification of metadata common to all resources as well as creation of metadata and content specific to particular types of resources. Metadata are split into science and system metadata and are packaged in a format compatible with existing data repositories, such as DataOne. HydroShare resource types include different data types used in the hydrology community and also models and workflows. Another fundamental concept in HydroShare is that of a Party, representing a person or organization that may be a resource creator or contributor, or a system user or group with access privileges for resources. This paper describes the resource centric approach being used, presents the data models for resources and parties, and discusses the application programming interface functions used as a foundational interface layer in the architecture of the system for accessing resources.

INTRODUCTION

Hydrologic information is collected by many individuals and organizations in government and academia for many purposes, including ambient monitoring of the water environment and specific investigations of hydrologic processes and environments. It is thus dispersed and heterogeneous. Advancing understanding in hydrology requires discovery of, access to, and integration of data and information from multiple sources. It requires integrated modeling to codify and synthesize knowledge in a form that is testable through reproducible comparisons to data. It requires collaboration. Data and modeling information technology systems, or cyberinfrastructure, are required to address these problems and enhance the ability of hydrologic scientists to collaborate by sharing data and models. HydroShare is an open source system being developed to meet these needs. HydroShare will provide a community collaboration web site that enables users to easily discover and access data and models, retrieve them to a desktop computer or perform analyses in a distributed computing environment that includes grid, cloud, or high performance computing model instances as necessary. We envision that HydroShare will enable more rapid advances in hydrologic understanding through collaborative data sharing, analysis, and modeling. Understanding will be advanced through the ability to integrate information from multiple sources. Outcomes (data, results, models) can then be published as new resources that can be shared with collaborators.

The Consortium of Universities for the Advancement of Hydrologic Science Inc. (CUAHSI) develops community infrastructure and services to advance hydrologic science. The CUAHSI Hydrologic Information System (HIS) [1] is a services-oriented-architecture established to support the sharing of hydrologic data. It is comprised of hydrologic databases and servers connected through web services as well as software for data publication, discovery and access that provides mechanisms for publishing, cataloging, discovering and accessing information using standardized web services. HIS supports the storing of point observations data using the Observations Data Model (ODM) [2]; sharing data through well-defined web services using a HydroServer [3, 4]; and the discovery and integration of this information through the open source HydroDesktop client [5]. HydroShare expands the data sharing capability of the CUAHSI HIS by broadening the classes of data accommodated, expanding capability to include the sharing of models and model components, and taking advantage of emerging social media functionality to enhance information about and collaboration around hydrologic data and models.

A HydroShare resource is the fundamental unit of content within the system. HydroShare manages resources as social objects that can be shared, annotated, endorsed, collaborated around, etc. HydroShare resources are information packages that are described by basic metadata. A resource can be a dataset in any of a set of supported (known) formats, a model, or a generic file or set of files whose structure and format may be unknown to HydroShare but that a user wants to group and share.

In this resource-centric approach (Figure 1), tools or models may perform actions (analyses) on resources of known types and formats that are stored in the resource store. This provides one way for models to interact. The output from one model can become the input to another model. The same analysis and visualization tools can be used on data from multiple models. This interaction among models and data with the resource store as the exchange point allows a separation of roles and supports compartmentalization of focus and re-use of best practice methods. For example, a modeler can focus on modeling while taking advantage of standardized best practice analysis, visualization, calibration, loading and discovery tools. An analyst can consider information from multiple models without having to be an expert in the details of each model.

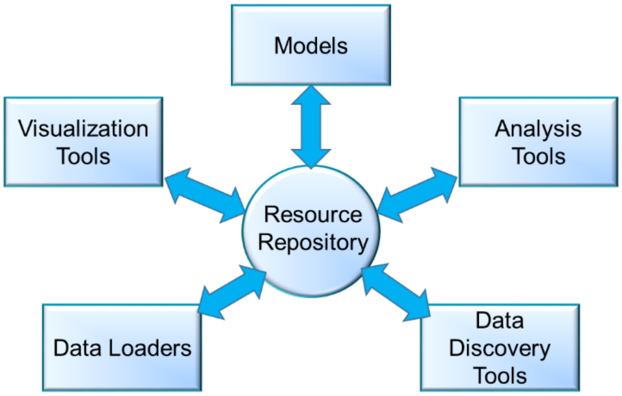

Figure 1. Resource Repository Centric Paradigm for Modeling and Analysis.

The way that data is structured can enhance or inhibit the analysis it can support and can also determine the level of interoperability with other cyberinfrastructure. The data model for representing resources is thus critical, and we describe the resource data model we are using to structure the representation of data, models, and other content in the system. Additionally, HydroShare is intended to have a social component. We also describe the data model for parties in the system. Party here refers to a person or organization that may be a resource creator or contributor, or a HydroShare user or group who collaborate on resources. We discuss the access control scheme developed to balance simplicity and transparency with the need to allow users to manage who has access to their resources. HydroShare is a web application. The architecture separates the web interface from the underlying functionality through a web service application programming interface (API) designed to encapsulate as much of the functionality as possible to support interoperability with other systems. This API is another foundational element of the system that is discussed here. HydroShare is at an early stage of development. We conclude with remarks on how we anticipate that the functionality being developed will contribute to advances in hydrology through collaboration.

RESOURCE DATA MODEL

In HydroShare, all content is represented using a Resource Data Model that separates system and science metadata and has elements common to all resources as well as elements specific to individual resource types. The HydroShare Resource Data Model was designed largely based upon the Open Archives Initiative's Object Reuse and Exchange (OAI-ORE) standard [6], which is a standard for the description and exchange of aggregations of web resources. HydroShare uses the BagIt File Packaging Format [7], which has been used in several library and digital curation implementations. BagIt is a hierarchical file packaging format designed specifically for disk-based storage and transfer of digital content. The HydroShare Resource Data Model was also heavily influence by the way DataONE [8] represents datasets (DataONE also uses OAI-ORE and BagIt), and HydroShare's adoption of these technologies promotes compatibility with DataONE (one goal of the HydroShare system). The following are essential properties of HydroShare resources:

- Resources may be comprised of a single content file or an aggregation of multiple content files.
- Resources containing multiple content files may have a hierarchical file/directory structure.
- Each resource is described by "science" metadata, which is a separate unit of digital content that details the properties of the resource (i.e., resource level metadata).
- Each content file within a resource may be separately described by a "science" metadata document that is considered to be part of the resource content (i.e., content level metadata).
- Each resource is accompanied by "system" metadata that contains system level attributes of the resource, including time stamps, ownership, access control rules, etc.
- Persistent identifiers, access control, versioning, sharing, and cataloging for discovery are all managed at the resource level in HydroShare.
- Resources may have additional external identifiers (e.g. Digital Object Identifiers, or DOIs) that uniquely identify the resource but are added after a resource is initially created.

We intend to have HydroShare support a broad class of resource types designed to include the data types and formats commonly used in hydrologic research. These include time series, geographic features and raster (gridded data), multidimensional space-time data sets, and composite resources to represent complex datasets such as river geometry. Models and their associated files are simply regarded as another type of resource that can be shared and manipulated within HydroShare. The heterogeneity in file types, formats, and potential hierarchical structure of content required a file-based data model and drove the selection of technologies used by the Resource Data model. Development is now underway to add functionality for these various resource types to HydroShare. Specifying a HydroShare resource type requires definition of:

- Data content, structure, and format
- The name and type of all data and metadata elements
- Which metadata elements are required or optional
- Which metadata elements are from a vocabulary
- File formats for import, storage, and export

Adding a resource type to HydroShare required development of type specific tools that enable users to open, visualize, convert, analyze, and otherwise manipulate the contents of resources beyond standard create, read, update and delete and social interaction functionality of generic resources. HydroShare does not prevent users from uploading resources containing objects that are unknown to the system, but does not provide any value added functionality for those resources other than allowing users to upload them, describe them with metadata, set access control permissions on them, share them, comment upon and rate them, and download them.

PARTY MODEL

In HydroShare a party refers to a person or organization that may be a resource creator or contributor, or a HydroShare user or group who collaborate on resources. The representation of parties in HydroShare has been developed drawing from ORCID, Friend of a Friend (FOAF), and the ScienceBase user model [9-11]. The HydroShare party model is shown in Figure 2 depicting party as a top level class with person and organization as subclasses. The model records associations between persons and organizations. Scholar is a further subclass of person to represent a user with an account in the system. Scholar is used as a general term rather than the system specific HydroShare user. ScholarGroup is an aggregation of scholars.

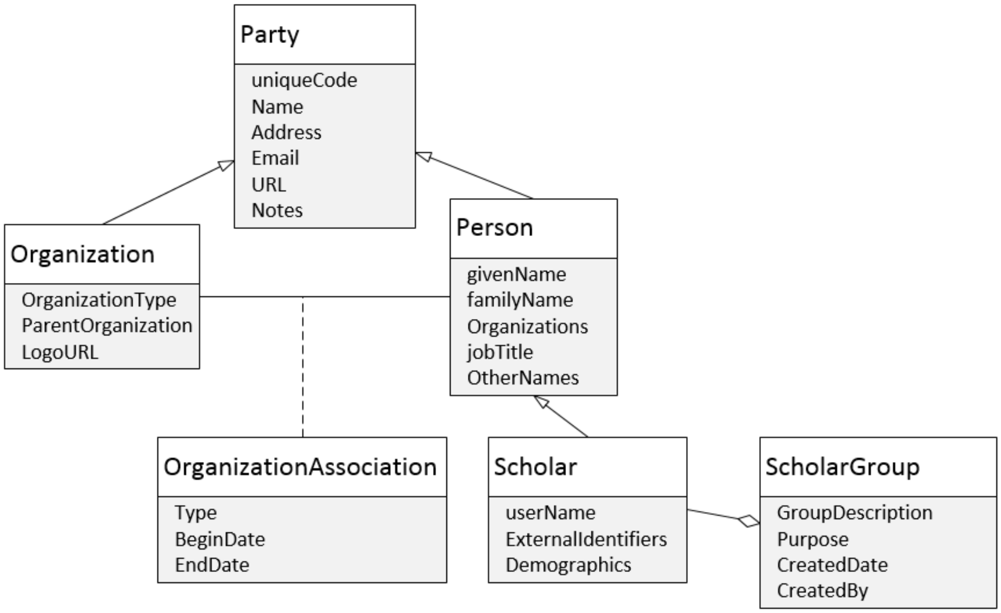

Figure 2. HydroShare Party Model

Dublin core metadata for each resource includes creator and contributor elements that record authorship (intellectual responsibility) and contributions to the content of resources. These elements are represented using the Party concept so that a resource creator or contributor may be a person or organization. Note that a resource creator or contributor does not have to be a HydroShare user (scholar), although users and groups may be resource creators or contributors. Organizations and users have associated types allowing for identification of community attributes (University, Commercial, Personal, etc) and demographic information that needs to be tracked. A person can be associated with more than one organization, and they can list previous associations with time. A user can list multiple external identities, such as Research Identifiers, which allow for disambiguation of users, and for HydroShare to interact with external systems. The Access Control rules (below) apply to system users and groups that may be different from resource creators and contributors.

ACCESS CONTROL

Drawing upon experience with and analysis of the rules used by other systems (e.g., Windows, Unix, DropBox, Google, Facebook) the access control functionality conceived for HydroShare strives for a system that is simple and straightforward, with the mindset of eliminating unnecessary mechanisms in favor of a set of behaviors specific to the tasks involved in data-centered research. The purpose for these rules is to balance the protection of and sharing of content in HydroShare. We want to make it easy for users to share content in HydroShare with whom they want at the level of access they choose to facilitate collaboration and publication. Access control is managed at the level of single resources and may be assigned to users or groups, or opened to the public. There are four access control levels:

- **View:** can see, download, and otherwise view the resource. Can comment upon and rate the resource.
- **Change:** can edit the resource and/or its scientific metadata. Cannot change system metadata, including sharing attributes (see below). Can also View the resource.
- **Full:** can change anything about the resource, including who has Change and View privileges, subject to restrictions on changes for formally published resources. Can delete the resource if not published. Can change the owner.
- **Owner:** functionally that can do everything that is associated with Full, but carries the burden of having the resource count against his/her quota.

The following non-mutually exclusive sharing attributes may be set for each resource in HydroShare:

- **Do not distribute -** Entities who receive access to the resource may not grant that access to others.
- **Discoverable -** The metadata is publically accessible independent of the ability to access data.
- **Public -** The resource is accessible to the public. Setting Public automatically sets Discoverable, but not the other way round.
- **Published -** The resource has been permanently published in the system. It is assigned a permanent identifier and is made immutable.

"Do not distribute" is intended to give resource owners control over who the resource is distributed to, a use case that is deemed important by some in the community to protect the intellectual content of resources that may contain unpublished data. "Discoverable" enables users to make resource metadata public while protecting the resource content. "Public" makes the resource accessible to everyone (i.e., users not signed in or without accounts). "Published" is reserved for resources published in a formal capacity, e.g. published with a DOI. These sharing attributes are held at the resource level and are separate from the settings assigned to users (View, Change, Full, Owner).

Once a resource is formally published the concept of change access control is removed. In addition the changes that a user with Full and Owner privileges can make are reduced to a limited set of operations on the metadata such as adding "is referenced by" and "is part of" references without changing the data. All users inherit view privileges for published resources and may continue to comment on and rate them. Also, do not distribute and discoverability restrictions are removed for published resources.

APPLICATION PROGRAMMING INTERFACE

HydroShare has web service Application Programming Interfaces (APIs) that support interaction between client applications and the main HydroShare system and facilitate development of client applications to interact with HydroShare resources. As a design principle, the HydroShare web service APIs are exposing the same functionality that can be accomplished through the HydroShare web user interface so that client applications can mimic that functionality (Figure 3). In general, HydroShare web services are being implemented using a Representational State Transfer (REST) based approach using HTTP as the transport protocol and XML and/or JSON for encoding messages. The current development approach is investigating use of the Tastypie (http://tastypieapi.org/) web service API framework for the Django (https://www.djangoproject.com/) web framework. A comprehensive list of API functions categorized into Resource Management, User Management and Authorization, Resource Discovery, and Social Functionality has been specified and is under development. To the extent possible, calling parameters are being replicated between a Python Client API, Web Service interface API, and an Internal Python Server API to promote consistency in development and use.

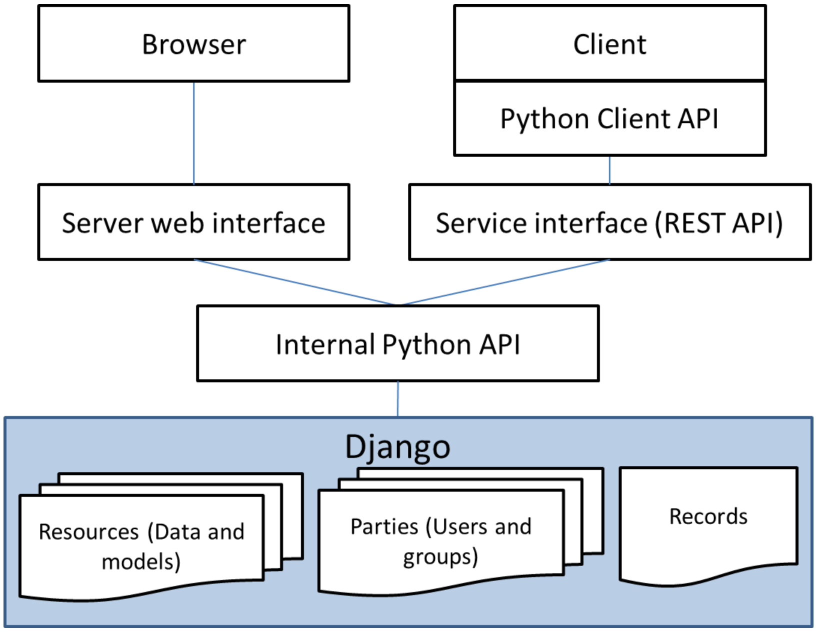

Figure 3. HydroShare high level architecture with Application Programming Interfaces

CONCLUSIONS

Much of the functionality described here is still in an early stage of development, with limited functionality available through our beta system. We envision, once this functionality is more fully developed, that HydroShare will enable more rapid advances in hydrologic understanding through collaborative data sharing, analysis, and modeling. The resource-centric design presented here for HydroShare will serve as the foundation for a community collaboration web site that enables users to easily discover and access data and models, retrieve them to a desktop computer or perform analyses in a distributed computing environment that includes grid, cloud, or high performance computing model instances as necessary. The flexible Resource Data Model we have designed for HydroShare supports a diverse class of resource types, focusing on management of resources as social objects within the system. The social and collaborative functionality of HydroShare will promote the development and sharing of data, models and best practice tools for data processing and analysis. The enhanced ability to collaborate, integrate information from multiple sources, and share and comment on research outcomes (data, results, models) as new resources will make it easier to build on the results of others, advance the information content and value of shared data and model resources, and ultimately increase the rate at which advances are made in hydrologic understanding. The provenance and metadata that HydroShare maintains will enhance reproducibility and transparency of the research conducted using HydroShare resources.

ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation under collaborative grants OCI-1148453 and OCI-1148090. 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.

REFERENCES

[1] Tarboton DG, Horsburgh JS, Maidment DR, Whiteaker T, Zaslavsky I, Piasecki M, et al., editors. Development of a Community Hydrologic Information System. 18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation; 2009 July: Modelling and Simulation Society of Australia and New Zealand and International Association for Mathematics and Computers in Simulation; (2009).

[2] Horsburgh JS, Tarboton DG, Maidment DR, Zaslavsky I. A Relational Model for Environmental and Water Resources Data. Water Resour Res. (2008);44:W05406.

[3] Horsburgh JS, Tarboton DG, Schreuders KAT, Maidment DR, Zaslavsky I, Valentine D, editors. Hydroserver: A Platform for Publishing Space-Time Hydrologic Datasets. 2010 AWRA Spring Specialty Conference Geographic Information Systems (GIS) and Water Resources VI; 2010; Orlando Florida: American Water Resources Association, Middleburg, Virginia, TPS-10-1; (2010).

[4] Conner LG, Ames DP, Gill RA. HydroServer Lite as an open source solution for archiving and sharing environmental data for independent university labs. Ecological Informatics. (2013);18(0):171-7.

[5] Ames DP, Horsburgh JS, Cao Y, Kadlec J, Whiteaker T, Valentine D. HydroDesktop: Web services-based software for hydrologic data discovery, download, visualization, and analysis. Environmental Modelling & Software. (2012);37:146-56.

[6] Lagoze C, Van de Sompel H, Johnston P, Nelson M, Sanderson R, Warner S. Open Archives Initiative Object Reuse and Exchange: ORE User Guide – Primer (2008) [10/23/2012]. http://www.openarchives.org/ore/1.0/primer.

[7] Boyko A, Kunze J, Littman J, Madden L, Vargas B. The BagIt File Packaging Format (v0.97) Network Working Group Internet Draft (2012) [2/6/2014]. http://tools.ietf.org/html/draft-kunze-bagit-10.

[8] DataONE. Data Observation Network for Earth. (2014) [04/07/2014]. http://www.dataone.org/.

[9] ORCID. ORCID XML (2014). http://support.orcid.org/knowledgebase/topics/32832-orcid-xml.

[10] Brickley D, Miller L. Friend of a Friend Vocabulary Specification 0.99 (2014). http://xmlns.com/foaf/spec/.

[11] USGS. ScienceBase Directory Services (2014). https://my.usgs.gov/confluence/display/sciencebase/ScienceBase+Directory+Services.