Jun 16th, 9:00 AM - 10:20 AM

Metadata for Describing Water Models

Mohamed M. Morsy
University of Virginia, mmm4dh@virginia.edu

Jonathan L. Goodall
University of Virginia, goodall@virginia.edu

Christina Bandaragoda
University of Washington, cband@uw.edu

Anthony M. Castronova
Utah State University, tony.castronova@usu.edu

Jane Greenberg
University of North Carolina, janeg@email.unc.edu

International Environmental Modelling and Software Society (iEMSs)
7th Intl. Congress on Env. Modelling and SoftwareSan Diego, California, USA,
Daniel P. Ames, Nigel W. T. Quinn, and Andrea E. Rizzoli (Eds.)
http://www.iemss.org/society/index.php/iemss-2014-proceedings

Metadata for Describing Water Models

Mohamed M. Morsy [a], **Jonathan L. Goodall**[a], **Christina Bandaragoda** [b] **Anthony M. Castronova**[c] **and Jane Greenberg** [d]

[a]*Department of Civil and Environmental Engineering, University of Virginia, PO Box 400742, Charlottesville, VA, USA (mmm4dh@virginia.edu, goodall@virginia.edu)*

[b]*Department of Civil and Environmental Engineering, University of Washington, 201 More Hall, Box 352700, Seattle, WA 98195-2700 (cband@uw.edu)*

[c]*Utah Water Research Laboratory, Utah State University, 8200 Old Main Hill, Logan, UT 84322-8200 (tony.castronova@usu.edu)*

[d]*School of Information and Library Science, University of North Carolina, 216 Lenoir Drive, CB #3360, 100 Manning Hall, Chapel Hill, NC, USA 27599-3360 (janeg@email.unc.edu)*

Abstract: Computer models are widely used in hydrology and water resources management. A large variety of models exist, each tailored to address specific challenges related to hydrologic science and water resources management. When scientists and engineers apply one of these models to address a specific question, they must devote significant effort to set up, calibrate, and evaluate that model instance built for some place and time. In many cases, there is a benefit to sharing these computer models and associated datasets with the broader scientific community. Core to model reuse in any context is metadata describing the model. A standardized metadata framework applicable across models will foster interoperability and encourage reuse and sharing of existing resources. This paper reports on the development of a metadata framework for describing water models. We discuss steps taken to achieve this goal for the HydroShare system and within the context of a use case that describes a team-based hydrologic modeling project.

Keywords: Hydrologic modeling; Model metadata; Dublin Core Metadata Initiative; Dublin Core Application Profile

1 INTRODUCTION

Computer models are powerful tools for understanding hydrologic and water resource systems. Dozens of models have been built for simulating various aspects of physical, biological, chemical, geological, economic or social systems that influence or are influenced by hydrologic and water resource systems [Singh and Woolhiser, 2002]. These models have grown in complexity and simulate detailed processes happening within water systems. Set up, calibration, and evaluation of the models now requires sophisticated supporting applications including GIS-based user interfaces, calibration tools, visualization software, and workflow systems to assist in the modeling process. There have also been recent efforts to advance the concept of integrated environmental modeling focusing on interdisciplinary challenges and the need to create modeling frameworks [Laniak et al., 2013].

This large and growing ecosystem of models and supporting applications has reached a point of complexity that necessitates methods and tools for organizing existing resources such as data inputs, outputs, and software code. Metadata, such as the date a model was created, the person or group that created the model, the model's geographic and temporal coverage, and other attributes, are required to support discovery and reuse of the model. In some cases, these metadata can be extracted from journal or conference papers, however only documenting this information within papers is not ideal because the information is not structured as standard, machine-readable metadata. New systems are needed and are being built to better capture this modeling metadata. One such system is HydroShare being developed by a team of researchers in collaboration with the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) with support from the US National Science Foundation (NSF).

HydroShare is envisioned as a web-based collaborative system with the goal of sharing, accessing, and discovering hydrologic data and models [Tarboton et al., 2013]. Users will be able to create accounts in HydroShare and upload their own datasets, models, modeling tools, model applications, and related resources. Users have the ability to share their resources with only certain selected users or make their resources publicly accessible. The goal of HydroShare is to advance hydrologic science by enabling the scientific community to more easily and freely share the tools of their trade: not just the scientific publication summarizing a study, but also the data and models used to create the scientific publication.

One of the key steps in building a model sharing system like HydroShare is clearly defining resources stored within the system, their metadata, and their interrelationships. While HydroShare is envisioned as a system that can be used to share many types of resources, the focus of this paper is on model resources. Many of the concepts discussed in the paper are applicable to model sharing systems in any discipline, although our focus is in the hydrology and water resources discipline. The objective of this paper is to define the metadata for describing a model resource within the context of HydroShare. This is presented in two parts: first we describe elements of a foundation domain model for modeling concepts and second we apply the domain model concepts for a use case. The longer-term goal of the work documented in this paper aims to create a Dublin Core Application Profile (DCAP) for hydrologic modeling concepts [Coyle and Baker, 2009]. DCAP is a generic construction used to design metadata records to meet the needs of a specific application where new metadata terms can be used in addition to or rather than the ones defined by the Dublin Core Metadata Initiative (DCMI) (see http://dublincore.org/documents/profile-guidelines). Our goal is to align with existing metadata standards where advances have been made and ensure interoperability. The DCAP permits this type of development in the context of linked data and reuse of metadata properties.

2 Model Concepts and Relationships

Computational modeling in hydrology involves multiple tools that are used collectively to create model inputs, run models, and analyze model outputs. Figure 1 provides a basic structure for relating three key concepts in hydrologic modeling through Resource Description Framework (RDF) triples. RDF triples take the form subject-predicate-object for describing attributes of a resource (see http://www.w3.org/RDF). For example, a model program (subject) uses (predicate) model inputs (object). The figure depicts relationships between Model Program, Model Input, and Model Output resource types. The Model Program is the executable file containing the model's logic which sometimes is called the model engine. This Model Program generates Model Output, and Model Output is derived from Model Input. Not shown in this figure are metadata for each resource, which could include Dublin Core metadata properties (creator, contributor, description, etc.) and may be extended with additional metadata properties as well for particular resources types, as will be shown in the use case example. The figure is intentionally simple, but it can easily be expanded to include relationships with other resources. For example, there could be a 'is generated by' relationship between a graphical user interface (GUI) tool used to generate the model input files.

Another increasingly common approach involves modeling frameworks that allow for "plug-and-play" coupling of model components. Examples of these systems include the Community Surface Dynamics Modeling System (CSDMS) [Peckham et al., 2013], and modeling frameworks that adopt the Open Modeling Interface (OpenMI) standard [Gregersen et al., 2007; Castronova et al., 2013]. Figure 2 captures the key modeling concepts for plug-and-play modeling frameworks. A Model Framework is used to link Model Components. At the same time, a Model Framework is used to generate and

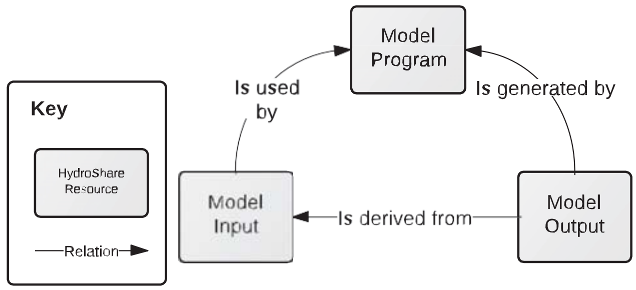

Figure 1. Describing relationships between three resource types, Model Input, Model Program, and Model Output, as RDF triples.

execute a Linked Model Configuration which contains the Linked Model Components. When a Linked Model Configuration is executed, it generates Model Output files by using Model Input files. Similar to Figure 1, the Model Output files are derived from the Model Input files. Again, this figure omits the additional Dublin Core metadata that would exist for each component. More extensive metadata could be captured for classes within this basic framework. For example, Elag and Goodall [2013] present a more detailed ontological framework for describing model components that could be used if addition model metadata is required.

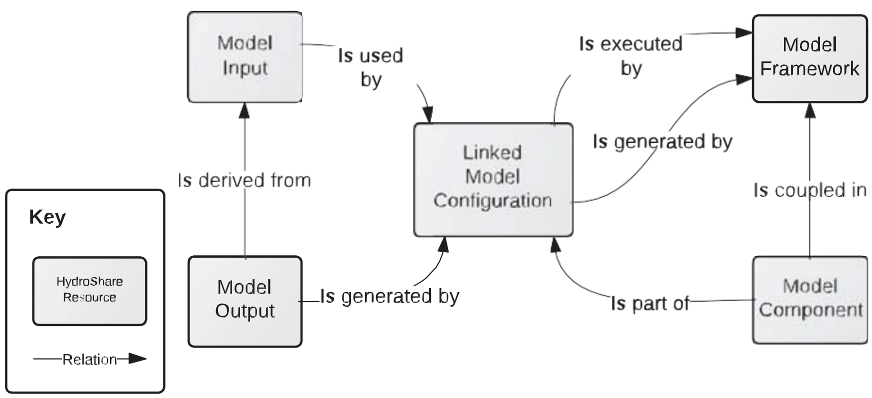

Figure 2. Describing the relationships between the resource types of a Linked Model Configuration that consists of Model Components coupled in a Model Framework. Like the Model Program concept, the Linked Model Configuration uses Model Input and generates Model Output.

In addition to the concepts introduced through the two examples of a Model Program and a Model Framework, there are other modeling concepts used within the hydrologic science community including Model Module, Model Script, and Model Workflow. A model module is source code written to expand the functionality of a model program to include additional physical processes. The model module concept is different from the model component concept in that the former is typically integrated into a model program through tight coupling approaches, as demonstrated in the example use case section. A model script is source code that is run using an interpreter. Examples of this are scripts written in languages such as Python, Matlab, R, and Perl. These scripts might be used to transform a data set into a model input file as part of a preprocessing step. There is also the concept of a model workflow that is used for creating model inputs with a common example being a workflow of geoprocessing routines used to delineate subbasins from a digital elevation model.

3 EXAMPLE USE CASE

Consider the following use case as an example application that provides a complicated, although not uncommon, scenario for team-based hydrologic modeling. The goal of this work is to design a way to capture the metadata content of this use case so that it is machine understandable and generalizable to similar use cases of other models being used by other teams over time to conduct hydrologic science investigations.

Use Case: Christina Bandaragoda, the University of Washington (UW) Ecohydrology research associate, would like to use HydroShare for coordinating, documenting, and sharing the various model and data resources generated through the following research project. A multi-organizational team will use the Distributed Hydrology Soil Vegetation Model (DHSVM), which is an open source model program created at UW by Mark Wigmosta under the supervision of Dennis Lettenmaier in 1994 [Wigmosta et al., 1994, 2002], in the Nooksack River Basin, Washington, USA for a hydrologic study. The study will include model runs using DHSVM. The study will make use of a water quality module developed by Ning Sun, also under the supervision of Dr. Lettenmaier, that was recently added to the DHSVM model program [Sun et al., Prep]. A new glacier module created by Gary Clarke [Jarosch et al., 2013] will be integrated into DHSVM for the study by PhD candidate Chris Frans, advised by Erkan Instanbulluoglu and a member of the UW Ecohydrology Research Group. The modeling work will use observational data at stations managed by Oliver Grah, Water Resources Program Manager at the Nooksack Indian Tribe. Climate projections and model forcing data will be developed by UW Land Surface Hydrology Group research scientist Matt Stumbaugh. Geospatial model inputs will be developed by Western Washington University graduate student advised by Robert Mitchell.

Figure 3 uses RDF relationships to capture the basic resources described in this use case and their interrelationships. The use case describes a model package that is a composite resource consisting of a model program, model input, and model output resources. These model input and output resources include datasets created by different individuals. The model program is also a composite resource that includes module resources created by different individuals.

Additional metadata can be added to more fully capture the use case (Figure 4). For example, each resource has a title, creator, and other metadata that follows the Dublin Core metadata standard. In some cases, it would be necessary to go beyond these core and very general metadata attributes to include domain specific attributes. For example in Figure 4 a person has an academic advisor and a module has an integrator that was responsible for integrating the module into a model program.

Each resource identified within the use case would be a distinct resource within HydroShare. The distinct resources in HydroShare would be interrelated so that, for example, it is possible to trace relationship such as the Model Input resource “was created by” a particular person. That person may be a registered HydroShare user, in which case the resource could be link to the person’s user profile including their affiliations, contact information, and a list of resources the person has created or contributed to within HydroShare. Resource creators can better follow how their resources are being used and have a means for quantifying the impact of the work that went into creating resources.

4 SUMMARY, DISCUSSION, AND NEXT STEPS

Creating a framework for describing the metadata for water models is a complicated process and here we present the first steps along that path. While we have presented a domain model for the concepts of a model program and a modeling framework, this domain model needs to be further expanded for other hydrologic modeling concepts. Nonetheless, it is promising that with relatively few extensions beyond the Dublin Core metadata standard, it is possible to capture key metadata concepts for a fairly

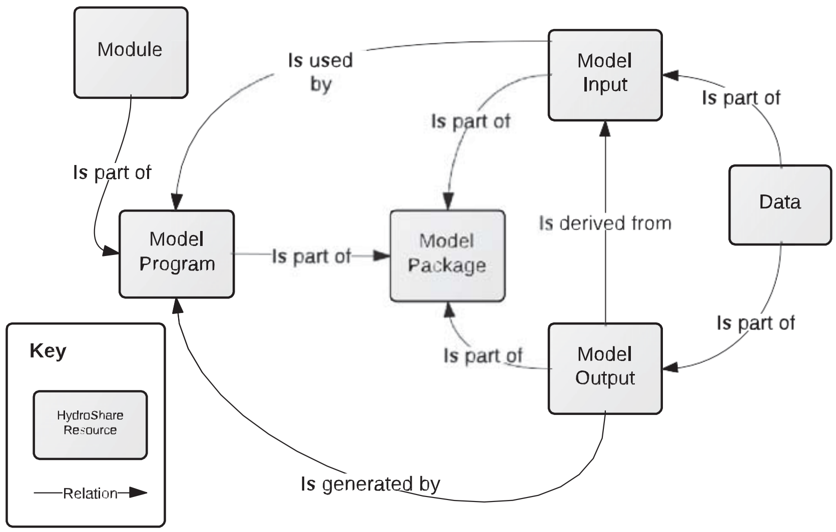

Figure 3. A basic framework for describing the example use case that consists of the resource types of a Model Module, Model Program, Model Package, Model Input, Model Output, and Data.

complex use case of team-based hydrologic modeling. With additional use cases of similar complexity, additional concepts, attributes, and relationships will become apparent and it is expected that the metadata framework begun here will need to be expanded to include these additional features.

The implications of designing a functional metadata framework for use in hydrologic sciences are far-reaching, as it impacts the ability of researchers to conduct efficient searches for available resources and the efficiency of building cross-domain-organization-disciplinary collaborations. This is a challenge in many fields of science where multiple physical processes are linked in space and time. The proposed metadata framework designed for use in HydroShare is an opportunity for exploring broader applicability. By working with World Wide Web Consortium (W3C), linked data/Semantic Web standards, such as RDF, and using the Dublin Core metadata standards as a foundations, the model resource work presented in this paper will have applicability across other communities gathering scientific data and models.

Metadata modeling advances, to date, have generally focused on means for packaging component parts of digital resources, as seen with the Functional Requirements for Bibliographic Records (FRBR) [Standing Committee of the IFLA Section on Cataloguing, 1997], and Metadata Encoding and Transmission Standard [METS, 2014]. These developments have advanced metadata practices mainly within the scholarly communications venue, and fall short in the area of scientific modeling. Conceptualizing, developing, and implementing a metadata framework for modeling resources within the HydroShare domain that is not at odds with these other developments will allow for greater interoperability and sustainability, and provide a framework that is of value to many other disciplines and domains that rely heavily on modeling.

The work reported in this paper underscores a long term goal for the CUAHSI HydroShare project. We have focused on the metadata for these resources as one piece of the larger problem. Additional work designing processes and implementing designs as software tools is still required. For example, work is needed to define file contents and directory structures for model resources so that the software system can extract information from the uploaded model resource. Work on this and similar questions is underway within the larger HydroShare team because the process will be similar for any resource,

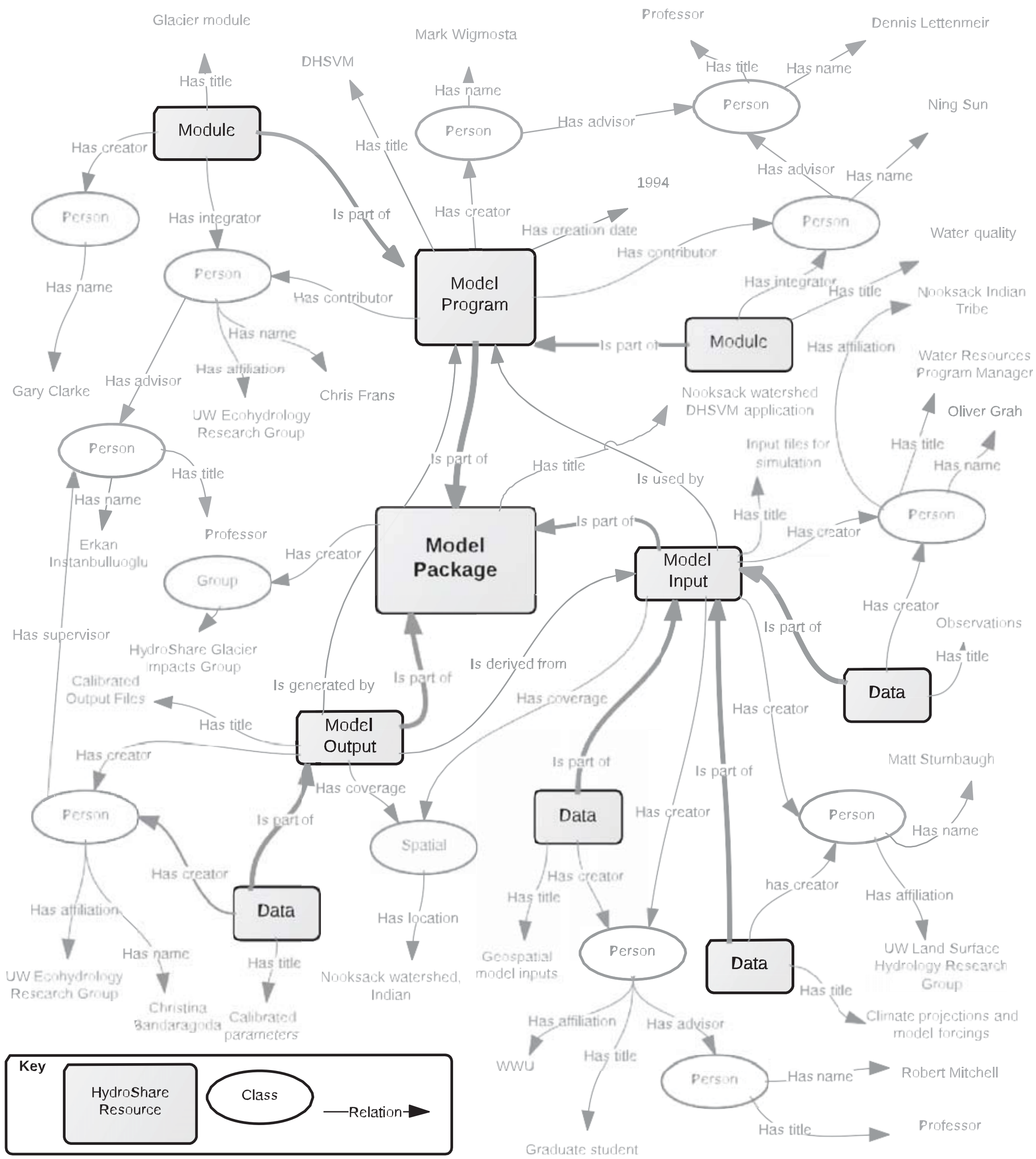

Figure 4. Expanding on Figure 3 to include include specific attributes described in the example use case. These additional attributes are shown in grey to highlight the core HydroShare resource objects within the figure.

not only model resources, managed within HydroShare [Tarboton et al., 2013].

Acknowledgments

This work was supported by the National Science Foundation under the award numbers 0940841 and 1148453. We acknowledge input from other team members of these projects, known as the CUAHSI HydroShare team and the DataNet Federation Consortium (DFC) team.

References

Castronova, A. M., Goodall, J. L., and Ercan, M. B. (2013). Integrated modeling within a hydrologic information system: an OpenMI based approach. *Environmental Modelling & Software*, 39:263–273.

Coyle, K. and Baker, T. (2009). Guidelines for Dublin Core application profiles. Dublin Core Metadata Initiative, http://dublincore.org/documents/profile-guidelines.

Elag, M. and Goodall, J. L. (2013). An ontology for component-based models of water resource systems. *Water Resources Research*, 49(8):5077–5091.

Gregersen, J., Gijsbers, P., and Westen, S. (2007). OpenMI: Open Modelling Interface. *Journal of Hydroinformatics*, 9(3):175–191.

Jarosch, A. H., Schoof, C. G., and Anslow, F. S. (2013). Restoring mass conservation to shallow ice flow models over complex terrain. *The Cryosphere*, 7:229–240.

Laniak, G. F., Olchin, G., Goodall, J., Voinov, A., Hill, M., Glynn, P., Whelan, G., Geller, G., Quinn, N., Blind, M., et al. (2013). Integrated environmental modeling: a vision and roadmap for the future. *Environmental Modelling & Software*, 39:3–23.

METS (2014). Metadata encoding and transmission standard (METS). Official Website: http://www.loc.gov/standards/mets.

Peckham, S. D., Hutton, E. W., and Norris, B. (2013). A component-based approach to integrated modeling in the geosciences: The design of CSDMS. *Computers & Geosciences*, 53:3–12.

Singh, V. P. and Woolhiser, D. A. (2002). Mathematical modeling of watershed hydrology. *Journal of hydrologic engineering*, 7(4):270–292.

Standing Committee of the IFLA Section on Cataloguing (1997). Functional requirements for bibliographic records: Final report. Approved by the Standing Committee of the IFLA Section on Cataloguing. (1997, amended and corrected through Feb. 2009).

Sun, N., Yearsley, J., and Lettenmaier, D. (In Prep). Effects of mid-twenty-first century climate and land cover change on the water quality of the puget sound basin, washington.

Tarboton, D. G., Idaszak, R., Horsburgh, J. S., Ames, D., Goodall, J. L., Band, L. E., Merwade, V., Couch, A., Arrigo, J., Hooper, R. P., Valentine, D. W., and Maidment, D. R. (2013). Hydroshare: An online, collaborative environment for the sharing of hydrologic data and models. Presented at the 2013 Fall Meeting, San Francisco, Calif., 9-13 Dec.

Wigmosta, M. S., Nijssen, B., Storck, P., Singh, V., Frevert, D., et al. (2002). *Mathematical models of small watershed hydrology and applications*, chapter The distributed hydrology soil vegetation model, pages 7–42. Water Resources Publications.

Wigmosta, M. S., Vail, L. W., and Lettenmaier, D. P. (1994). A distributed hydrology-vegetation model for complex terrain. *Water Resources Research*, 30(6):1665–1679.