Environmental Modelling & Software 92 (2017) 317–329

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

Evaluation of the OntoSoft Ontology for describing metadata for legacy hydrologic modeling software

Bakinam T. Essawy [a], Jonathan L. Goodall [a, *], Hao Xu [b], Yolanda Gil [c]

[a] Department of Civil and Environmental Engineering, University of Virginia, 351 McCormick Road, PO Box 400742, Charlottesville, VA 22908, United States
[b] Data Intensive Cyber Environment Center, University of North Carolina, Chapel Hill, NC, United States
[c] Information Sciences Institute and Department of Computer Science, University of Southern California, United States

ARTICLE INFO

Article history:
Received 23 November 2016
Accepted 30 January 2017

Keywords:
Hydrologic modeling
Scientific workflows
Metadata
Computational reproducibility

ABSTRACT

Metadata for hydrologic models is rarely organized in machine-readable forms. This lack of formal metadata is important because it limits the ability to catalog, identify, attribute, and understand unique model software; ultimately, it hinders the ability to reproduce past computational studies. Researchers have recently proposed an ontology for scientific software metadata called OntoSoft for addressing this problem. The objective of this research is to evaluate OntoSoft for organizing the metadata associated with a data pre-processing software workflow used in association with the Variable Infiltration Capacity (VIC) hydrologic model. This is accomplished by exploring what metadata are available from online resources and how this metadata aligns with the OntoSoft Ontology. The results suggest that past efforts to document this software resulted in capturing key model metadata in unstructured files that could be formalized into a machine-readable form using the OntoSoft Ontology.

1. Introduction

Hydrologists use many different computational models, with each model tailored to address specific questions and problems. Hydrological modeling has a long history, and many computational models have decades of development effort and many model versions behind them (Singh et al., 2002). In many cases, there has been splintering of the model code base where the original model code has started to be developed along different paths. This causes confusion as to which specific version of software was used for a given modeling application. Further complicating the issue, models often have supporting software beyond the physical process-representations within the model engine itself. This software is used to create input datasets for the model (i.e., data pre-processing) and to analyze or visualize the output from the model (i.e., data post-processing). Organizing and categorizing this broad collection of modeling software so that it is possible to uniquely identify the software used to perform a study is a significant challenge.

The need to better manage the growing volume of software used for hydrologic modeling is central to the larger challenge of computational reproducibility. The common approach for achieving reproducibility has been for researchers to provide sufficient detail within a journal paper's methods section to allow for reproducing the study's results. Growing complexity in computational analyses means this approach is no longer sufficient. Scientific disciplines are trying different approaches to address this problem including model repositories, documentation, on-line model execution, and scientific workflows (De Roure et al., 2009; Essawy et al., 2016; Gregersen et al., 2007; Lud et al., 2006; Roure et al., 2010). One of the main purposes of these approaches is to make models easier to reuse so that scientists can advance the model while achieving reproducibility and strengthening the decisions based upon these models (Cassey and Blackburn, 2006; Hutton et al., 2016; Scholten et al., 2000).

To achieve “reproducible software” (Peng, 2011; Gil et al., 2016a) for hydrologic modeling, not only does the software and data need to be shared, but also their associated metadata. Metadata is structured information for describing and explaining a digital resource that makes it easier to manage, retrieve, and use that resource (NISO, 2004). Metadata is now a common term for describing data sets, but metadata is less commonly used for describing software. Software for data collection, storage, retrieval,

* Corresponding author. University of Virginia, Department of Civil and Environmental Engineering, PO Box 400742, Charlottesville, VA 22904, United States.

E-mail address: goodall@virginia.edu (J.L. Goodall).

http://dx.doi.org/10.1016/j.envsoft.2017.01.024
1364-8152/© 2017 Elsevier Ltd. All rights reserved.

processing, and management has improved greatly, and has significantly contributed to the development of comprehensive distributed hydrological models (Singh et al., 2002). Capturing metadata for hydrologic modeling software is one of the steps required to make the software reproducible (Higgins, 2007; Mcdougal et al., 2016). Little attention has been paid to metadata for describing these software advances. Computational reproducibility also requires other advanced uses of standard software practices beyond metadata tools including version control, strong commenting and documentation, and code modularity.

The limited past efforts to define metadata for hydrologic models have largely focused on describing well maintained and widely used hydrologic models as a single information resource. Like data, however, there is a long-tail of software used to perform and support hydrologic modeling (Heidorn, 2008). Models are often the combination of smaller software modules or components contributed over time by a large number of individuals and groups. Taking a more granular view of models by diving into the details of the software provenance and attempting to capture this provenance using metadata is necessary for many reasons. Some of these reasons include 1) providing attribution for software contributions, 2) maintaining and archiving existing models, 3) providing information that aids in installing and executing models, and 4) ultimately fostering reproducibility.

Metadata for hydrologic models is being collected and recorded, but it is unstructured, informal and distributed. The available metadata for these models are scattered across model documentation, source code repositories, model publication repositories, user forums, and other publically available resources. Metadata such as who created the model, when the model was created, and the type of input and output data for the model can be found from these sources for many scientific models, but are not provided in human-readable form. Not having this information in a machine-readable form limits its utility and does not scale well to the growing volume of scientific software. Metadata needs to be in machine readable formats to be most useful (e.g. RDF, XML).

Efforts to establish more formalized, machine-readable formats for hydrologic model metadata include efforts through the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) HydroShare project and the Community Surface Dynamics Modeling System (CSDMS) project. HydroShare describes metadata for two key modeling concepts: a model program and a model instance. The model program is the software for executing the model and the model instance is the input files required for executing the model (Horsburgh et al., 2015; Morsy et al., 2014; Tarboton et al., 2014). A metadata framework has been proposed for both of these concepts that extend the Dublin Core Metadata Standard. The CSDMS project created a catalog of model programs across the surface dynamics community, which includes hydrology, and captured metadata for these model programs (Peckham and Goodall, 2013; Peckham et al., 2013).

Recent related activities have focused on designing standard metadata for describing software with a particular focus on scientific software. OntoSoft is a project that is part of the National Science Foundation EarthCube Initiative and provides an ontology and portal for addressing the challenge of capturing metadata for scientific software in a formal way (Gil et al., 2016b, 2015). The metadata captured by the OntoSoft Ontology focuses on the knowledge needed for software sharing and reuse (Ratnakar et al., 2015). It is recommended for documenting software in scientific papers that follow best practices for reproducible research, open science, and digital scholarship (David et al., 2016; Gil et al., 2016a), and has been used to document scientific software in published articles, e.g., (Fulweiler et al., 2016; Pope, 2016; Yu et al., 2016). OntoSoft is used in the research reported in this paper because it was designed and developed by experts in the semantic metadata community, in contrast to past efforts for hydrologic model metadata that was designed and developed by hydrologists. An underlying question that the research reported in this paper begins to address is whether this more general scientific metadata ontology is appropriate and useful for describing hydrologic modeling software.

The objective of this study is to advance prior efforts for formalizing model metadata in hydrology by evaluating the OntoSoft Ontology as a means for structuring model metadata. The evaluation is performed using a data pre-processing workflow for the Variable Infiltration Capacity (VIC) hydrologic model that consists of multiple software components written by different individuals over time. The VIC model is used by a large community; over 500 publications used this model since 1993. The analysis begins by exploring what metadata hydrologists have already captured in unstructured forms. It then shows how this metadata could be organized into structured, machine-readable metadata using the OntoSoft Ontology. Therefore, the primary contribution of this work is an evaluation of the OntoSoft Ontology for describing software relevant to hydrologic modeling. This is done by first understanding what metadata for hydrologic modeling software are already embedded in online resources, and then testing how this metadata maps to the OntoSoft Ontology.

2. Background

2.1. Variable Infiltration Capacity (VIC) model pre-processing workflow

VIC is a macro scale hydrologic model that applies water and energy balances to simulate terrestrial hydrology at a regional spatial scale (Liang et al., 1996). Like many hydrologic models, the VIC model requires significant effort to prepare its input data. Fig. 1 shows the data processing workflow used to generate the meteorological and land surface input datasets for a VIC model simulation. This workflow consists of a sequence of 15 data processing steps, each step requiring input datasets from different sources, and many of the datasets having unique data models (Billah et al., 2016). These scripts are written with different programming languages including Fortran 77, C, and C++. Shell scripts are used throughout the workflow to execute these steps and perform other commands required to complete the data processing tasks.

The workflow is divided into four categories as shown in Fig. 1. The first category of scripts process the precipitation and the air temperature datasets, the second category of scripts process the land surface datasets including topography, soil, and vegetation data, the third category of scripts process the wind speed dataset, and the last category of scripts create the final model input files for meteorological datasets. The datasets processed by the workflow are shown as ovals and include 1) meteorological forcing files (i.e., precipitation, wind, and minimum and maximum air temperature), 2) soil and vegetation parameter files, and 3) basin geospatial files. The primary inputs for the workflow are shown as parallelograms and include datasets from 1) the National Oceanic and Atmospheric Administration (NOAA) National Climatic Data Center (NCDC) (now the National Centers for Environmental Information (NCEI)), 2) the National Center for Atmospheric Research (NCAR) National Centers for Environmental Prediction (NCEP), 3) the National Aeronautics and Space Administration (NASA) Land Data Assimilation System (LDAS), 4) the United States Geological Survey (USGS) HYDRO1K dataset, and 5) the PRISM Climate Group PRISM dataset.

This work addresses the challenges of creating metadata for the individual scripts within the VIC data processing workflow shown in Fig. 1. A significant amount of work by other scientists has gone