



OntoSoft: A Community Software Commons for the Geosciences

<http://www.ontosoft.org>

The Context

The OntoSoft project is part of the National Science Foundation's EarthCube Initiative. The NSF EarthCube Initiative aims to enable scientists to solve challenging problems that span diverse geoscience domains. This requires not only data sharing, it requires new forms of knowledge sharing. The focus of OntoSoft is to promote knowledge sharing about the software developed for geosciences.

The Challenges

Geosciences software embodies important scientific knowledge that should be explicitly captured, curated, managed, and disseminated. Recurring issues of provenance and uncertainty of data could be better addressed with improved treatment of geoscience software: it is easy to see that nothing describes data more precisely than the software that generates or uses the data. Software is implicitly linked to its documentation, datasets it uses or produces, publications, and ultimately scientific theories. Scientists recognize the value of sharing software to avoid replicating effort or to reproduce results from others. However, the stewardship of software in geosciences must be greatly improved.

First, while modeling frameworks have dramatically improved software sharing, there are orders of magnitude more codes devoted to preparing data for input to a model (what we could call "pre-model software") and preparing data that results from a model ("post-model" software). Studies show that scientists spend between 60%-80% of a project's effort collecting and preparing data before doing new science. This would indicate a significant overhead in developing pre-model software for data preparation is only rarely shared (e.g., through libraries such as NetCDF-Java and NumPy) and rarely reused, particularly across disciplines. This leads to wasted investments, particularly for younger researchers that are often charged with such tasks, which often go unnoticed.

Second, the inaccessibility of software as an explicit science product that, like data, should be shared and reused leads to barriers for those who cannot afford such investments, and to barriers for software experts that could otherwise become involved in supporting software efforts.

Finally, the education of future scientists crucially depends on the ability to actively explore problems by experimenting with science-grade data and software. This will not be possible unless science software is captured and properly disseminated. In summary, although geoscientists program a lot of code to analyze their data, that important software is often not shared and rarely preserved.

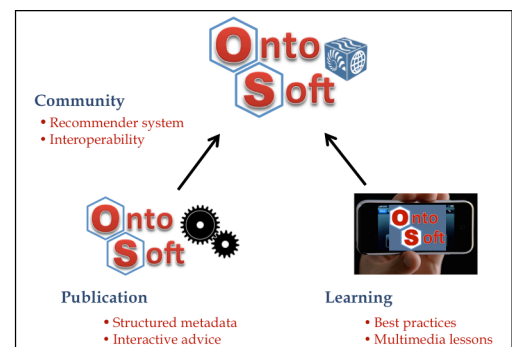
Our Approach

The goal of OntoSoft is to enable the creation of a germinal ecosystem for software stewardship in geosciences that will empower scientists to manage their software as valuable scientific assets in an open transparent mode that enables broader access to that software by other scientists, software professionals, students, and decision makers.

Scientific software stewardship requires a combination of cyberinfrastructure, social infrastructure, and professional development infrastructure. Our cross-disciplinary team has had direct experience with a variety of aspects of the scientific software lifecycle, from conception to development, deployment, characterization, integration, composition, and dissemination through open source communities and geosciences modeling frameworks.

Our work focuses on:

1. **Facilitating software publication** through a personal assistant that guides a user through best practices. Users choose the degree of investment they are willing to make in componentizing, describing, licensing, and maintaining their software. It will assist in metadata capture for open source publication, the formation of communities around the software, and set up mechanisms for software citation and credit. This guidance is designed around explicitly articulated best practices for open software sharing and reuse.
2. **Enabling broad software dissemination** through a "software commons" for geosciences that supports software contributions (prepared through OntoSoft or otherwise), software discovery through multi-faceted search, and that fosters social interactions through dynamic formation of communities of interest. OntoSoft interoperates with existing software repositories and modeling frameworks in geosciences by importing descriptions of their existing content and in turn advertising to those frameworks new contributions. OntoSoft is based on explicit incentives from social sciences and open source communities.



3. **Providing just-in-time training materials** through an annotated collection of educational units (videos, screen captures, decision trees, reports) ranging from basic education to professional training on all aspects of software stewardship. These materials will run the gamut from simple to complex, e.g., from suggesting the use of the most popular open source license for beginners to walking through a decision tree of nuanced choices for advanced users. These materials will be seamlessly integrated with the OntoSoft user interfaces, and present opportunities for contextualized learning as needed in the context of a user's context of interaction with the framework.

Initial Focus

Our focus to date has been on the OntoSoft portal that to facilitate software publication. OntoSoft provides:

- Intelligent assistance to describe software: how to use it appropriately, what kinds of data, how it relates to other software, and other important metadata that helps promote reuse
- Sophisticated search capabilities to find software for specific needs, based on license or type of data produced
- Interactive advice on software sharing, including open source publication, forming successful developer communities, and other software sharing topics

PIHM	PIHMgis	DrEICH	TauDEM	WBMsed
Identify, Describe, Use, Share	Identify, Describe, Use, Share	Identify, Describe, Use, Share	Identify, Describe, Use, Share	Identify, Describe, Use, Share
Is there any test data available for the software?	Is there any test data available for the software?	Is there any test data available for the software?	Is there any test data available for the software?	Is there any test data available for the software?
Test Data Location: http://sourceforge.net/projects/pihm/	Test Data Location: http://online.library.wiley.com/doi/10.1002/2013WR015167/full	Test Data Location: http://rosdms.colorado.edu/wiki/Model:TauDEM#Testing	Test Data Location: http://rosdms.colorado.edu/wiki/Model:WBMsed#Testing	Test Data Location: http://rosdms.colorado.edu/wiki/Model:WBMsed#Testing
Test Data Description: Upper Juniata River 875 km ² ; see: http://sourceforge.net/projects/pihm/model/	Test Data Description: Two test DEMs are included in the repository, both from Wayne National	Test Data Description: The Log an River DEM is a small test dataset useful for learning how to use the software	Test Data Description: Extensive input dataset is available on the CSDMS SP2C (search at: http://rosdms.colorado.edu/wiki/Model:WBMsed#Testing)	Test Data Description: Extensive input dataset is available on the CSDMS SP2C (search at: http://rosdms.colorado.edu/wiki/Model:WBMsed#Testing)
What are domain specific keywords for this software? (eg: hydrology, climate)	What are domain specific keywords for this software? (eg: hydrology, climate)	What are domain specific keywords for this software? (eg: hydrology, climate)	What are domain specific keywords for this software? (eg: hydrology, climate)	What are domain specific keywords for this software? (eg: hydrology, climate)
Basins, Continental	Basins, GIS	Geomorphology, Hydrological, Bedrock channel erosion	Hydrologically corrected DEM, Watershed	Sediment flux, Global model, Hydrological mode
What Operating Systems can the software run on?	What Operating Systems can the software run on?	What Operating Systems can the software run on?	What Operating Systems can the software run on?	What Operating Systems can the software run on?
Unix Windows Linux Mac OS	Unix Windows Linux Mac OS	Unix Linux	Unix Windows Linux Mac OS	Unix Linux

The technical underpinning of this portal is the OntoSoft ontology, designed to capture what a scientist would want to know about someone else's software in order to use it in their own work.

OntoSoft already interfaces with current model repositories such as CSDMS, and soon CIG and others.

Science Drivers

Our work is driven by demonstration scenarios in the context of the Critical Zone Observatory (CZO) project. Many of the new CZO models are extensions to existing models but their development is often not coordinated with the management and ongoing development of the model they are derived from. For example, during the first years of the Shale Hills CZO funding, several water and energy models were developed and now these models are being extended to model transport processes, stable isotopes, biogeochemistry and the carbon-nitrogen system. Our framework will provide connections that will provide significant benefit to this emerging, cross-disciplinary science.








We also actively gather science requirements through: 1) an Early Career Advisory Committee that is broad and diverse in its composition and cuts across geosciences disciplines, 2) a series of NSF-funded EarthCube end-user community workshops through 2013 and 2014 in order to generate a better understanding of the cyber-infrastructure needs of the broad geoscience community, 3) EarthCube Research Coordination Networks (RCNs) established to foster focused science community activities, and 4) the EarthCube Science Standing Committee and Technical and Architecture Standing Committee.

Benefits

The OntoSoft project will result in a germinal social site for the EarthCube, where scientists can discover alternative approaches to release free software, use intelligent interfaces to explain how their software works, and form productive communities around software projects.

This research has the potential to fundamentally transform geosciences by making scientific software readily available to researchers and citizen scientists for efficient data analysis.

More broadly, this work will improve our understanding of how to promote software sharing in science, support better software stewardship, and capture metadata for scientific software.

						<p>OntoSoft is funded by the National Science Foundation under the EarthCube Initiative through grants ICER-1343800 and ICER-1440323.</p>   <p>For more information, contact us at ontosoft@gmail.com</p>
<p>Chris Duffy Civil Engineering Penn State U.</p>	<p>Yolanda Gil (PI) Intelligent Systems U. Southern California</p>	<p>James Howison Information Studies U. Texas Austin</p>	<p>Chris Mattmann Software Engineering U. Southern California</p>	<p>Scott Peckham Hydrology U. Colorado</p>	<p>Erin Robinson Virtual Community ESIP/FES</p>	