OntoSoft

Software Stewardship for the Geosciences

www.ontosoft.org

ICER-1343800
ICER-1440323

Christopher J. Duffy
Dept. of Civil and Environmental Engineering
Penn State University

Yolanda Gil
Information Sciences Institute and Department of Computer Science
University of Southern California

Chris Mattmann
NASA/Jet Propulsion Laboratory and Department of Computer Science
University of Southern California

Scott Peckham
Institute of Arctic and Alpine Research
University of Colorado Boulder

Erin Robinson
Foundation for Earth Sciences

Problems Addressed

Today, most software developed by scientists is never shared

- There are repositories of model software (e.g., CSDMS), also code repositories (e.g., GitHub)
- However, software is rarely shared, particularly all the data pre-processing and visualization software.

Problem: Hard to build on other work, especially across disciplines

- Software captures valuable geoscience knowledge and should be shared
- "*Scientists and engineers spend more than 60% of their time just preparing the data for model input or data-model comparison*" (NASA A40)

Open science

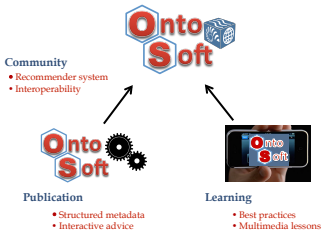
- Science products, including software, are of general societal importance [Holdren's OSTP memo, 2013] [NSF 2015]
- Reproducible science requires open software
- Open source software development is more sustainable

Scientists want to get credit for software much more than for data [Nature Metrics survey, 2010]

OntoSoft Project Goals

An on-line community for sharing knowledge about geosciences software

- Intelligent assistance to publish and describe new software: how to use it appropriately, what kinds of data, how it relates to other software, propagating provenance
- Recommender system and sophisticated search capabilities to find software that fits science needs
- Interactive advice on open source software, forming successful developer communities, and other software sharing topics



Novel research contributions

- An intelligent user interface that organizes the interaction in terms of science tasks (e.g., understand assumptions, do research with the software, cite the software, etc)
- An ontology to represent metadata for scientific software (http://www.ontosoft.org/software)
- Automated crawlers that extract metadata for the software from the user's web site (e.g., extracting license and other information from GitHub)
- Training modules to help scientists learn to share and describe software

Benefits

- Open science: Easy to disseminate models and software across disciplines
- Accelerate research: By reusing software
- Reproducibility: Easy to replicate results
- Accessibility: For non-programmers
- Quality: Best software that is well tested
- Integration: Well-described software is easier to integrate

OntoSoft: An Ontology for Software Metadata

Identify
Locate – unique identifier
has name (desc)
has short description (desc)
has software category (desc)
has unique ID (uniqueID)
has project web site (location+)

Understand
Relate – domain knowledge
has domain keywords (desc)
has uses and assumptions (desc)
has use limitations (desc)
similar software (desc)
Trust – quality and ratings
has creator (agent+)
has publisher (agent+)
has major contributor (agent+)
commitment of support (desc)
has adopters (entity+)
has use information (desc)
has use statistics (desc)
used in publication (citation+)
has benchmark information (desc)
has salient qualities (desc)
has funding sources (desc)
has rating (rating+)

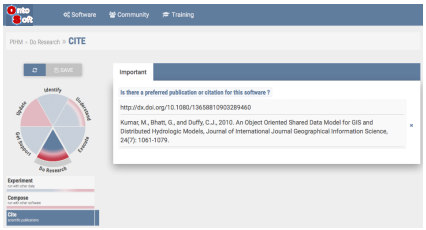
Do Research
Experiment – run with other data
has input (i-o)
has input parameter (i-o)
has output (i-o)
has relevant data sources (desc)
Compose – run with other software
has interoperable software (desc+)
has composition description (composition)
Cite – scientific publications
has preferred citation (citation+)

Execute
Access – download
has code location (location)
has executable location (location)
has license (license+)
Install – execution requirements
has documentation (location)
has installation instructions (desc)
has implementation language (language+)
has dependency (software version)
requires average memory (measurement)
supports operating system (os)
has average run time (desc)
has other implementation details (desc)
Run – testing execution
has test data (desc)
has test instructions (desc)

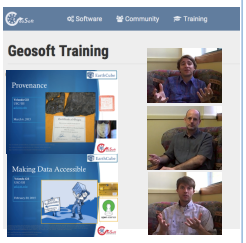
Get Support
Discuss – support and community
has email contact (email)
has software support (desc)

Update
Track – evolution
has software version (version)
has version release date (date)
supersedes (version)
superseded by (version)
Contribute – evolution
has active development (desc)
has software community (desc)

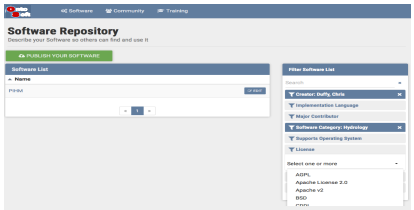
Publishing Software



Training



Finding and Reusing Software



Geoscience Papers of the Future