OntoSoft: A Distributed Semantic Registry for Scientific Software

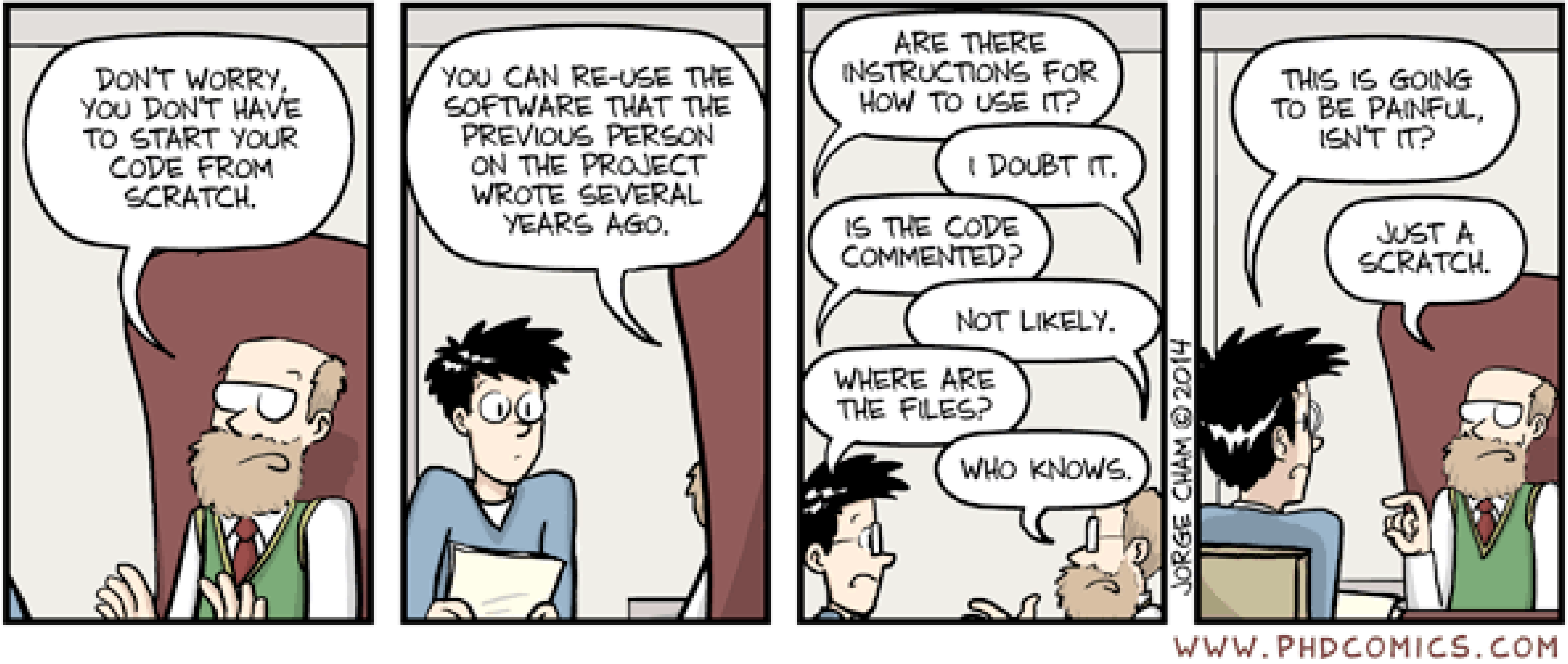
Yolanda Gil, **Daniel Garijo,** Saurabh Mishra, Varun Ratnakar

Information Sciences Institute
and Department of Computer Science
University of Southern California
@yolandagil, @dgarijov
{gil,dgarijo,saurabhm,varunr}@isi.edu

http://www.ontosoft.org

EarthCube Building Block

We have all been here…


DON'T WORRY, YOU DON'T HAVE TO START YOUR CODE FROM SCRATCH.
YOU CAN RE-USE THE SOFTWARE THAT THE PREVIOUS PERSON ON THE PROJECT WROTE SEVERAL YEARS AGO.
ARE THERE INSTRUCTIONS FOR HOW TO USE IT?
I DOUBT IT.
IS THE CODE COMMENTED?
NOT LIKELY.
WHERE ARE THE FILES?
WHO KNOWS.
JORGE CHAM © 2014
THIS IS GOING TO BE PAINFUL, ISN'T IT?
JUST A SCRATCH.
WWW.PHDCOMICS.COM

The Value of Software: Reproducibility

Illuminating the black box

nature

Note to biologists: submissions to *Nature* should contain complete descriptions of materials and reagents used.

Reporting Checklist For Life Sciences Articles

This checklist is used to ensure good reporting standards and to improve the reproducibility of published results. For more please read Reporting Life Sciences Research

Human lives

A Biostatistic Paper Alleges Potential Harm To Patients In Two Duke Clinical Studies

By Paul Goldberg

s aren't usually the place to go for sensational nt issue of the Annals of Applied Statistics is an exception.

A paper p may be harmed rely on biomar The paper iost of D ltip

COMPUTER SCIENCE

Accessible Reproduc

Science

ICYFORUM

putation in research grows, needed to expand recording,

Jill P. M

S rect. M tain a p edgeab

The New York Times

Science

NYTimes: Home - Site Index - Archive - Help

Reliability

Nobel Laureate Retracts Two Paper

新语丝Xin Yu Si (New Threads) Home Page

http://www.xys.org/

Most Visited Getting Started Latest Headlines Wolf-Haven Chicken Smoothie Howrse Motherload

Convert

Dartmouth and ... Aquaflow-work... Current Solicita... Description of ... 新语丝Xin Yu Si ...

新语丝

搜索新语丝

新语丝 New Threads

【一】【二】【三】【yahoogroups】

eb Feed(RSS): 【国内】【海外】

ed by PSI 留学生服务公司.

料，用手机浏览新语丝的方法

游牧虎 新语丝读者二春海边的家 预定热线：400-889-2006 www.yomovo.com

- 627名海内外知识分子关于肖传国派方舟子案的公开信
- 中国科学与学术诚信基金会呼吁为"扶助学术不端受害人专项基金"捐款
- 方舟子新书签名本邮购：国内、北美地区
- 网上购买：科学成就健康、批评中医、破解世界之谜、寻找生命的逻辑
- 网上购买：大象为什么不长毛 爱因斯坦信上帝吗？、带你走近科学、食品转基因
- 第十届"ＰＳＩ－新语丝"网络文学奖参选作品（257篇）

【肖传国案件】【丁肇诒案件】【刘荣杰事件】【杨玉圣事件】【其他专辑】【方舟子作品】【流氓专栏】【新语博客】【新语论坛】【中国青年报】【经济观察报】【科学新闻】【五柳村】【三思科学】【学术诚信基金】

Scientific integrity

The New York Times

Retracted Scientific Studies: A Growing List

3, 2015

Early report

Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children

RETRACTED

Friday, December 2, 2011 As of 12:00 AM New York 43° | 34°

THE WALL STREET JOURNAL. HEALTH

HEALTH INDUSTRY | DECEMBER 2, 2011

Scientists' Elusive Goal: Reproducing Study Results

bing how it had halted nearly two-thirds of experiments failed to match claims made

No Cure

When Bayer tried to replicate results of 67 studies published in academic journals, nearly two-thirds failed.

Source: Nature Reviews Drug Discove

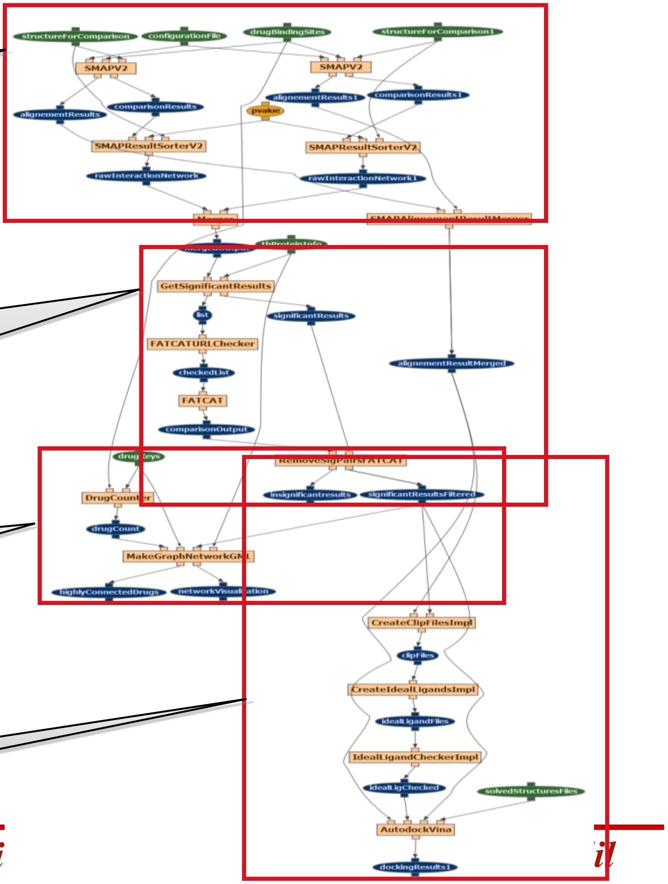
Financial

Trust

Quantifying the Value of Software through “Reproducibility Maps” [Bourne & Gil et al 12]

Work with P. Bourne of UCSD

- 2 months of effort in reproducing published method (in PLoS’ 10)
- Authors expertise was required



Comparison of ligand binding sites

Comparison of dissimilar protein structures

Graph network generation

Molecular Docking

Comparison of Ligand Binding Sites:

🟩 SMAP1	🟥 SMAP2	🟥 SMAP Result Sorter1	🟥 SMAP Result Sorter2	🟥 Merger	🟥 Align Result Merger	Minimal
🟩 SMAP1	🟩 SMAP2	🟩 SMAP Result Sorter1	🟩 SMAP Result Sorter2	🟩 Merger	🟩 Align Result Merger	Novice Author

Comparison of dissimilar protein structures:

🟥 GetSignificant Results	🟥 FATCAT URLChecker	🟥 FATCAT	🟥 Remove Significant Pairs	Minimal
🟥 GetSignificant Results	🟥 FATCAT URLChecker	🟩 FATCAT	🟥 Remove Significant Pairs	Novice
🟩 GetSignificant Results	🟩 FATCAT URLChecker	🟩 FATCAT	🟩 Remove Significant Pairs	Author

Docking

🟥 CreateClip Files	🟥 CreateIdeal Ligands	🟥 IdealLigand Checker	🟥 Autodock Vina	Minimal Novice
🟩 CreateClip Files	🟩 CreateIdeal Ligands	🟩 IdealLigand Checker	🟩 Autodock Vina	Author

Software Today

- There are repositories of domain specific software (e.g., geosciences)

CSDMS
COMMUNITY SURFACE DYNAMICS MODELING SYSTEM
CIG COMPUTATIONAL INFRASTRUCTURE for GEODYNAMICS
OSGeo
Your Open Source Compass
ESMF
Apache Open Climate Workbench

- There are general software repositories with no standard metadata

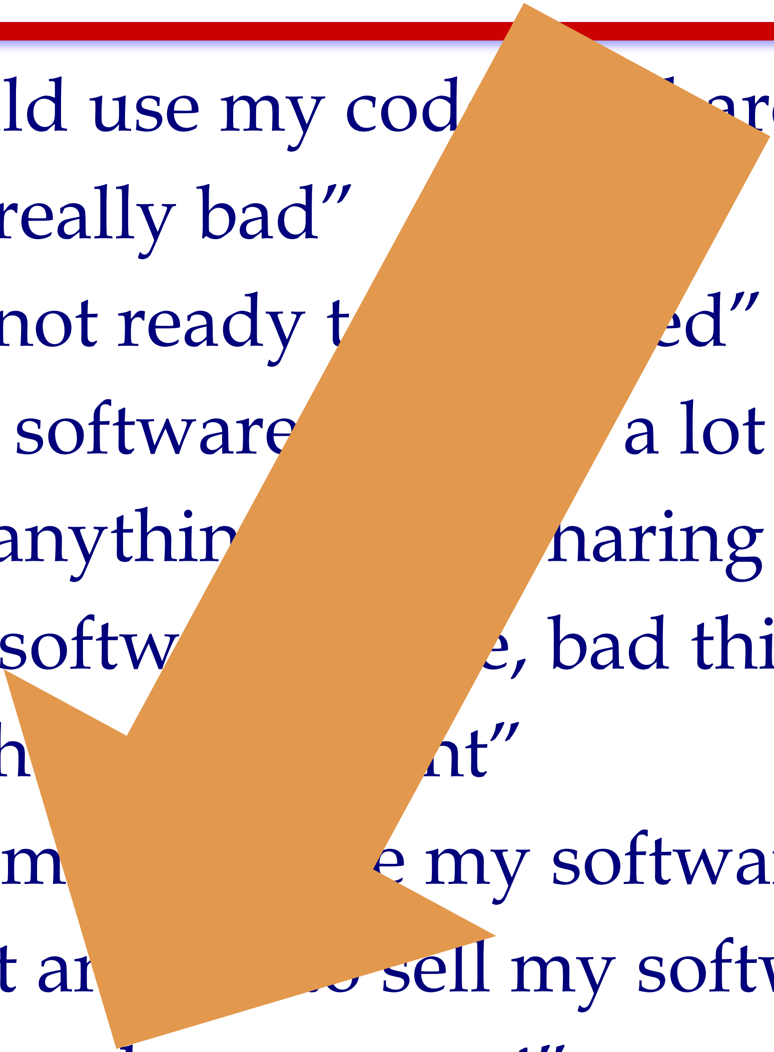
- Most scientists are not aware of the value of their software

“Dark Software”

- Models that are not published
  - Eg from a PhD thesis
- Data preparation software
  - Data pre-processing and QC can take up to 80% of a project’s effort
- Visualization software

“Dark Software” is the counterpart of “Dark Data” [Heidorn 2008]

Why Is Software Not Shared?

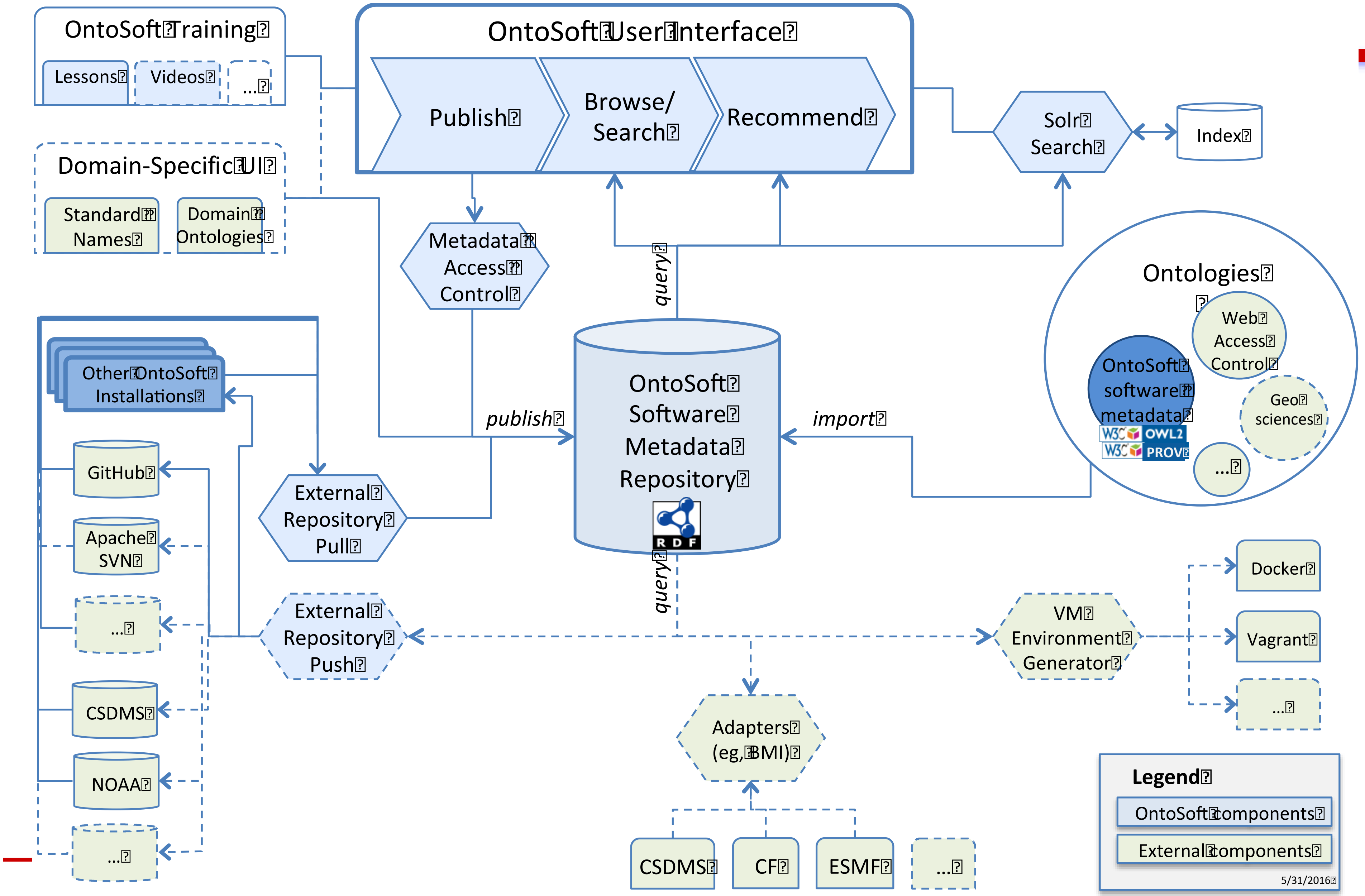


- “Noone would use my cod ared it”
- “My code is really bad”
- “My code is not ready t ed”
- “Sharing my software a lot of time”
- “I won’t get anythin haring my software”
- “I’ve shared softw e, bad things happened”
- “I work for th nt”
- “I want to com e my software”
- “I don’t want ar o sell my software”
- “I don’t know where to start!”

Contributions: OntoSoft

- Registry for software
  - Complements code repositories
  - Scientist-centered software metadata
  - Community curated software metadata
  - Training scientists on best practices

OntoSoft Architecture



5/31/2016

The OntoSoft Ontology for Describing Scientific Software Metadata [Gil et al 2015]

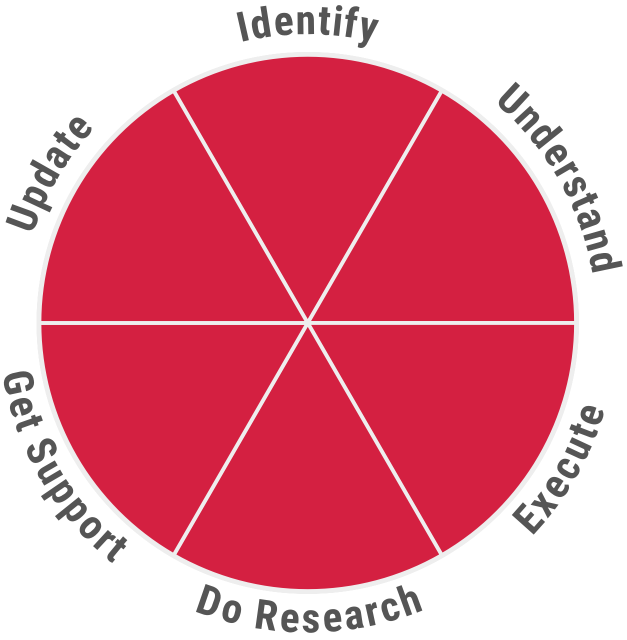
- **An ontology for scientific software metadata**
  - Intended to describe scientific software
  - **Designed with scientists in mind to guide them to deposit and describe their software in a software registry**

- **Major categories of metadata: what does a scientist need?**
  1. identify software
  2. understand what it does and its utility for research,
  3. execute the software,
  4. get support if questions arise,
  5. do research with it, and
  6. contribute to its development

http://www.ontosoft.org/software

OntoSoft Metadata Categories

Locate
unique description



Contribute
evolution

Track
versions

Discuss
support and community

Trust
quality and ratings

Relate
domain knowledge

Access
download

Install
execution requirements

Run
testing execution

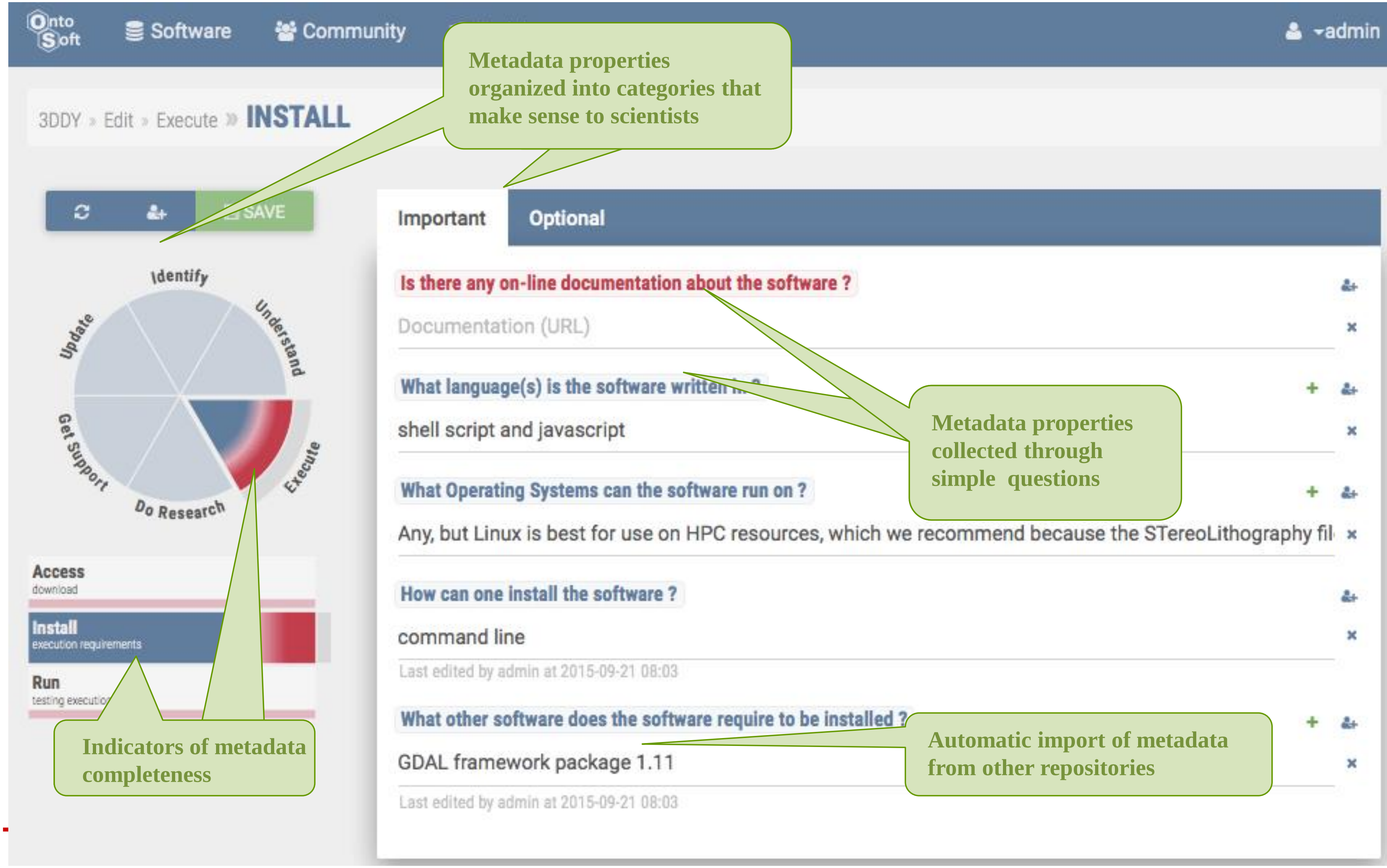
Experiment
run with other data

Compose
run with other software

Cite
scientific publications

http://www.ontosoft.org/portal

Describing Scientific Software in OntoSoft



3DDY » Edit » Execute » INSTALL

Access
download

Install
execution requirements

Run
testing executio

Important | Optional

Is there any on-line documentation about the software ?

Documentation (URL)

What language(s) is the software written in ?

shell script and javascript

What Operating Systems can the software run on ?

Any, but Linux is best for use on HPC resources, which we recommend because the STereoLithography fil

How can one install the software ?

command line

Last edited by admin at 2015-09-21 08:03

What other software does the software require to be installed ?

GDAL framework package 1.11

Last edited by admin at 2015-09-21 08:03

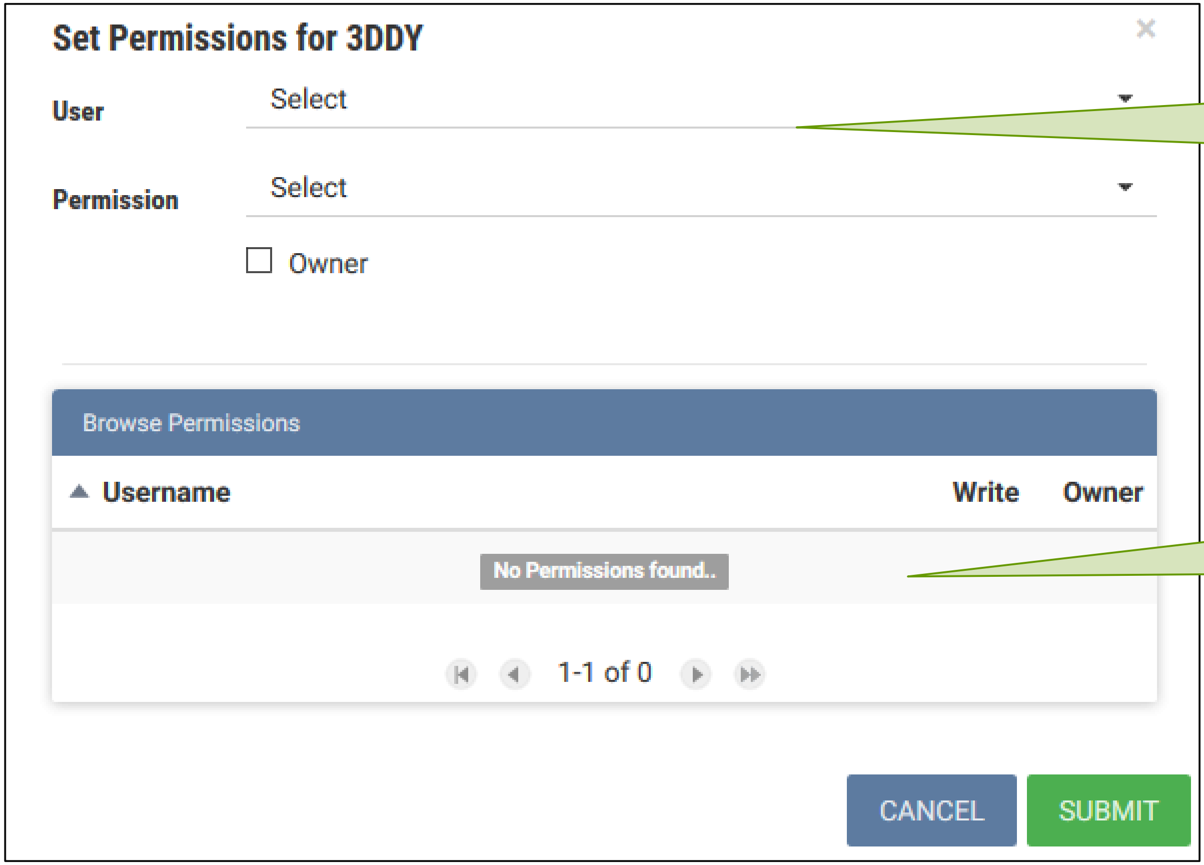
Metadata properties organized into categories that make sense to scientists

Metadata properties collected through simple questions

Indicators of metadata completeness

Automatic import of metadata from other repositories

Access control



Set Permissions for 3DDY ×

User Select ▾

Permission Select ▾

□ Owner

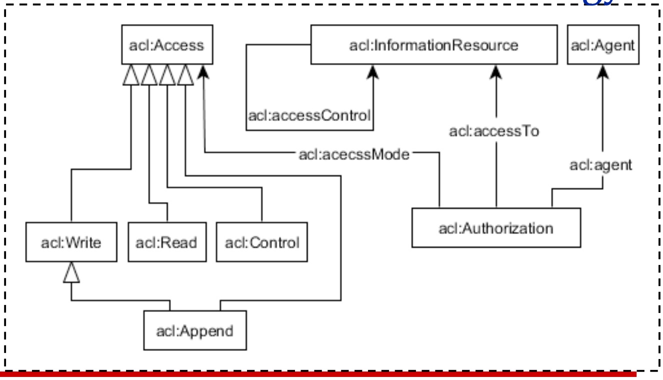
Browse Permissions

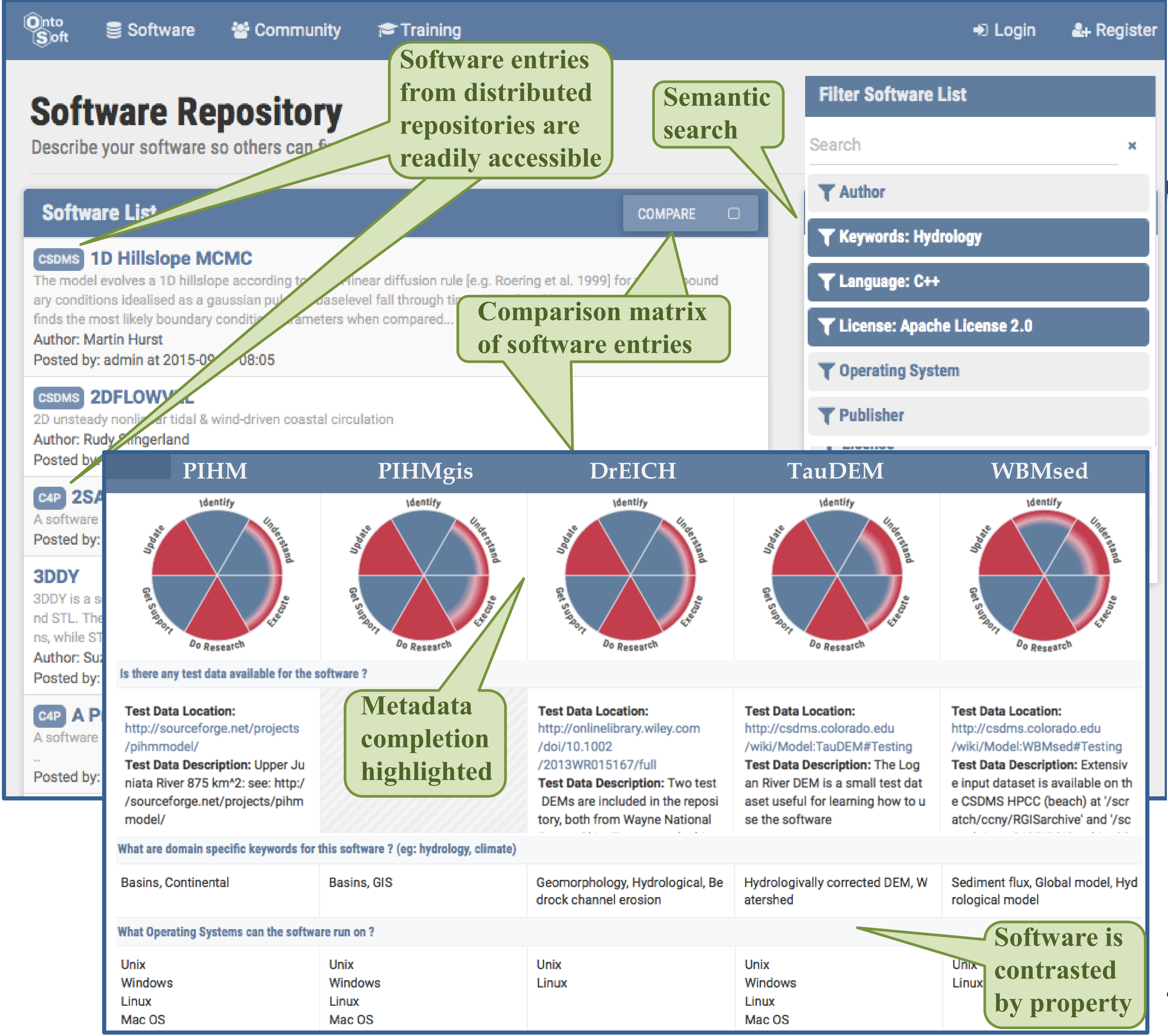
▲ Username	Write	Owner
No Permissions found..		

Setting permissions for editing 3DDY metadata

Users and permissions for the 3DDY software component

W3CWeb access control Ontology





Software Repository

Describe your software so others can f

Software List

CSDMS **1D Hillslope MCMC**
The model evolves a 1D hillslope according to ...linear diffusion rule [e.g. Roering et al. 1999] for ...bound ary conditions idealised as a gaussian pul... baselevel fall through ti... finds the most likely boundary conditi... rameters when compared...
Author: Martin Hurst
Posted by: admin at 2015-09 ... 08:05

CSDMS **2DFLOWV...L**
2D unsteady nonli...r tidal & wind-driven coastal circulation
Author: Rudy S...ngerland
Posted by

C4P **2SA**
A software
Posted by:

3DDY
3DDY is a s
nd STL. The
ns, while ST
Author: Su
Posted by:

C4P **A P**
A software
..
Posted by:

Filter Software List

Search ×

Author

Keywords: Hydrology

Language: C++

License: Apache License 2.0

Operating System

Publisher

License

	PIHM	PIHMgis	DrEICH	TauDEM	WBMsed
Is there any test data available for the software ?					
	Test Data Location: http://sourceforge.net/projects/pihmmodel/ **Test Data Description:** Upper Juniata River 875 km^2: see: http://sourceforge.net/projects/pihmmodel/		**Test Data Location:** http://onlinelibrary.wiley.com/doi/10.1002/2013WR015167/full **Test Data Description:** Two test DEMs are included in the repository, both from Wayne National	**Test Data Location:** http://csdms.colorado.edu/wiki/Model:TauDEM#Testing **Test Data Description:** The Logan River DEM is a small test dataset useful for learning how to use the software	**Test Data Location:** http://csdms.colorado.edu/wiki/Model:WBMsed#Testing **Test Data Description:** Extensive input dataset is available on the CSDMS HPCC (beach) at '/scratch/ccny/RGISarchive' and '/sc
What are domain specific keywords for this software ? (eg: hydrology, climate)					
	Basins, Continental	Basins, GIS	Geomorphology, Hydrological, Bedrock channel erosion	Hydrologivally corrected DEM, Watershed	Sediment flux, Global model, Hydrological model
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

Software entries from distributed repositories are readily accessible

Semantic search

Comparison matrix of software entries

Metadata completion highlighted

Software is contrasted by property

Collaborating with

SEN

C4P

EC3

Code meta initiative

Community

Early Career Advisory Board

Critical Zone Observatory

Omics

UK Software Institute

EarthCube RCNs

Publication

Learning

CSDMS

CIG

ESMF

EarthCube Building Blocks

FES/ ESIP

Software Carpentry

Conclusions

- Software is a valuable research product
  - Must embed best practices of software sharing into research activities
- Improve productivity, quality, reproducibility
- OntoSoft contributions
  - Ontology of scientific software metadata
  - Portal for software registry

http://www.ontosoft.org
http://www.ontosoft.org/software
http://www.ontosoft.org/portal

Do you want to use Ontosoft?
Let us know!

More Information

http://www.ontosoft.org
http://www.ontosoft.org/software
http://www.ontosoft.org/portal
http://www.ontosoft.org/gpf

- OntoSoft: Capturing Scientific Software Metadata. Yolanda Gil, Varun Ratnakar, and Daniel Garijo. *Proceedings of the Eighth ACM International Conference on Knowledge Capture (K-CAP)*, 2015.
- OntoSoft: A Distributed Semantic Registry for Scientific Software. Yolanda Gil, Daniel Garijo, Saurabh Mishra, and Varun Ratnakar. *Under review*, 2016.
- DRAT: An Unobtrusive, Scalable Approach to Large Scale Software License Analysis. Chris A. Mattmann, Ji-Hyun Oh, Tyler Palsulich, Lewis John McGibbney, Yolanda Gil, and Varun Ratnakar. *Proceedings of the Fourth International Workshop on Software Mining, held in conjunction with the 30th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, 2015.
- Cyber-Innovated Watershed Research at the Shale Hills Critical Zone Observatory. Xuan Yu, Chris Duffy, Yolanda Gil, Lorne Leonard, Gopal Bhatt, and Evan Thomas. *IEEE Systems Journal*, to appear.
- Collaborative Software Development Needs in Geosciences. Yolanda Gil, Eunyoung Moon and James Howison. *Proceedings of the Second Workshop on Sustainable Software for Science: Practice and Experiences (WSSSPE2), held in conjunction with the IEEE ACM International Conference on High Performance Computing (SC)*, New Orleans, LA, November 2014.
- Workflow Reuse in Practice: A Study of Neuroimaging Pipeline Users. Daniel Garijo, Oscar Corcho, Yolanda Gil, Meredith N. Braskie, Derrek Hibar, Xue Hua, Neda Jahanshad and, Paul Thompson and Arthur W. Toga. *Proceedings of the IEEE Conference on e-Science*, 2014.
- FragFlow: Automated Fragment Detection in Scientific Workflows. Daniel Garijo, Oscar Corcho, Yolanda Gil, Boris A. Gutman, Ivo D. Dinov, Paul Thompson and Arthur W. Toga. *Proceedings of the IEEE Conference on e-Science*, Guarujua, Brazil, October 2014.
- An Overview of Mobile Applications for Field Science. Anna Zeng, Kevin Zeng, Yolanda Gil, and Matty Mookerjee. *GeoSoft Project Report*, September 2014.
- The CSDMS Standard Names: Cross-Domain Naming Conventions for Describing Process Models, Data Sets and Their Associated Variables. Scott D. Peckham. *Proceedings of the Seventh International Congress on Environmental Modeling and Software*, San Diego, CA, June 2014.
- Web Applications that Share Level-12 HUC Data and Models of the CONUS. Lorne Leonard and Chris Duffy. *Proceedings of the Seventh International Congress on Environmental Modeling and Software*, San Diego, CA, June 2014.
- Intelligent Workflow Systems and Provenance-Aware Software. Yolanda Gil. *Proceedings of the Seventh International Congress on Environmental Modeling and Software*, San Diego, CA, June 2014.

EarthCube!

ICER-1440323
ICER-1343800

http://www.ontosoft.org
http://www.ontosoft.org/software
http://www.ontosoft.org/portal
http://www.ontosoft.org/gpf

Acknowledgements

- The OntoSoft project team includes Chris Duffy (PSU), Chris Mattmann (JPL), Scott Pechkam (CU), Ji-Hyun Oh (USC), Varun Ratnakar (USC), and Erin Robinson (ESIP)
- Thank you to James Howison (UT), Lisa Kempler (Matworks), and Greg Wilson (Software Carpentry) for their feedback on best practices for software sharing
- Thank you to the scientists and other colleagues that have contributed ideas and asked hard questions about software stewardship
- Thank you to the National Science Foundation and the EarthCube program for supporting this work