

Information Sciences Institute



USC Viterbi
School of Engineering

OntoSoft Tutorial

A Distributed Semantic Registry for Scientific Software

Daniel Garijo, Yolanda Gil

Information Sciences Institute and
Department of Computer Science

@dgarijov

dgarijo@isi.edu

Slides available: <https://goo.gl/J1pBqQ>

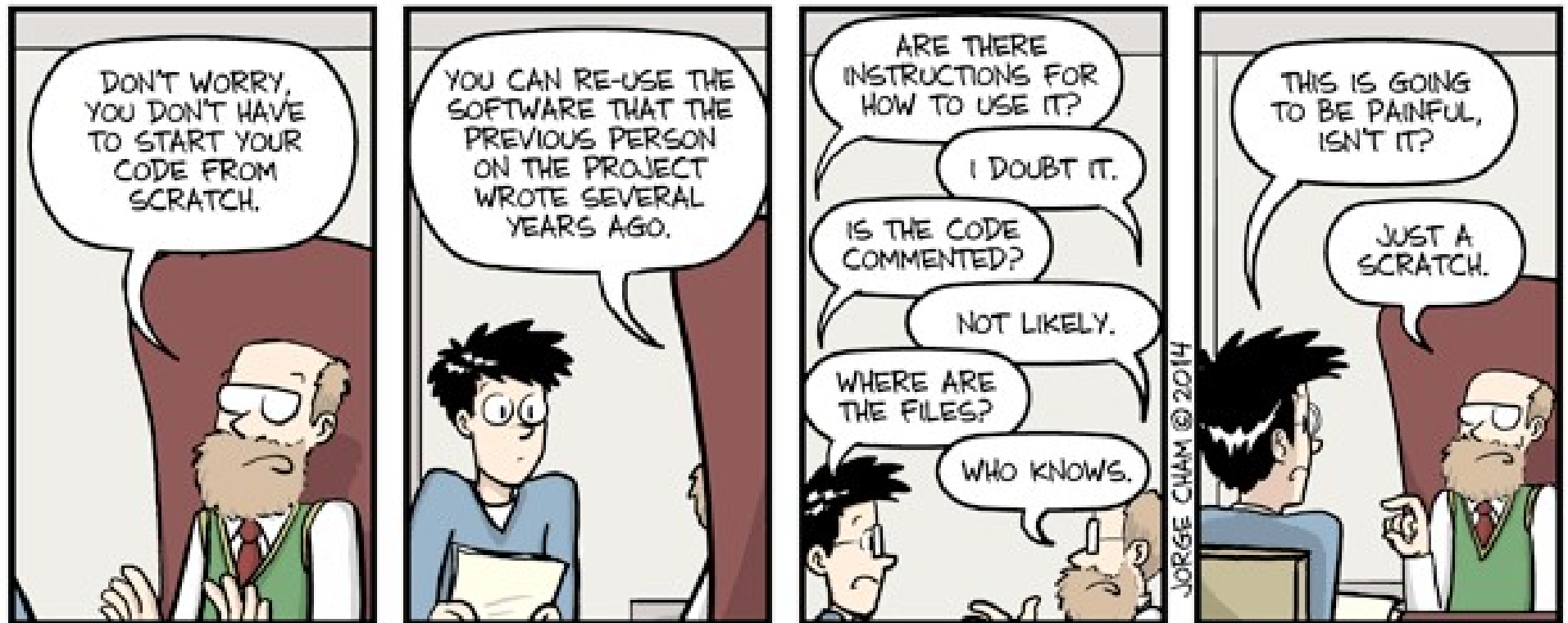


<http://www.ontosoft.org>



Information Sciences Institute

We have all been here...



WWW.PHDCOMICS.COM

The Value of Software: Reproducibility

Illuminating the black box

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

nature

Reporting Checklist For Life Sciences Articles

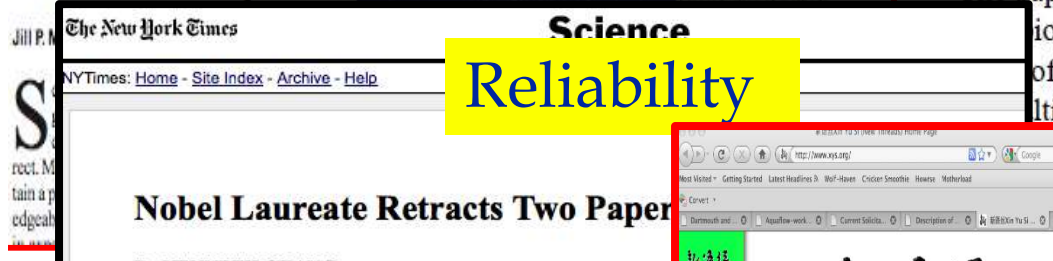
This checklist is used to ensure good reporting standards and to improve the reproducibility of published results. For more please read [Reporting Life Sciences Research](#).

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

By Paul Goldberg

Human lives

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



Reliability

The New York Times Retracted Scientific Studies: A Growing List



**Scientific
integrity**

exception.

A paper p
may be harmed
ely on biomar

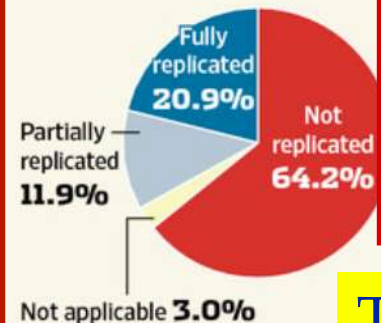
The paper



Financial

No Cure

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



Source: Nature Reviews Drug Discover

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

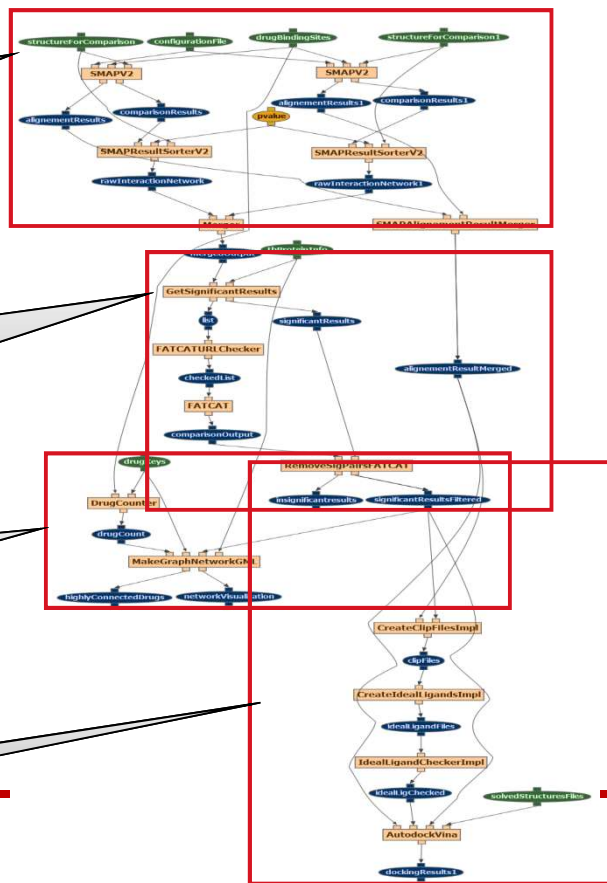
Comparison of ligand binding sites

Comparison of dissimilar protein structures

Graph network generation

Molecular Docking

Information Sciences Institute



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
SMAP1	SMAP2	SMAP Result Sorter1	SMAP Result Sorter2	Merger	Align Result Merger	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
CreateClip Files	CreateIdeal Ligands	IdealLigand Checker	Autodock Vina	Novice
CreateClip Files	CreateIdeal Ligands	IdealLigand Checker	Autodock Vina	Author

Software Today

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



Apache Open Climate Workbench

- There are general software repositories with no standard metadata (or curation process)



sourceforge

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

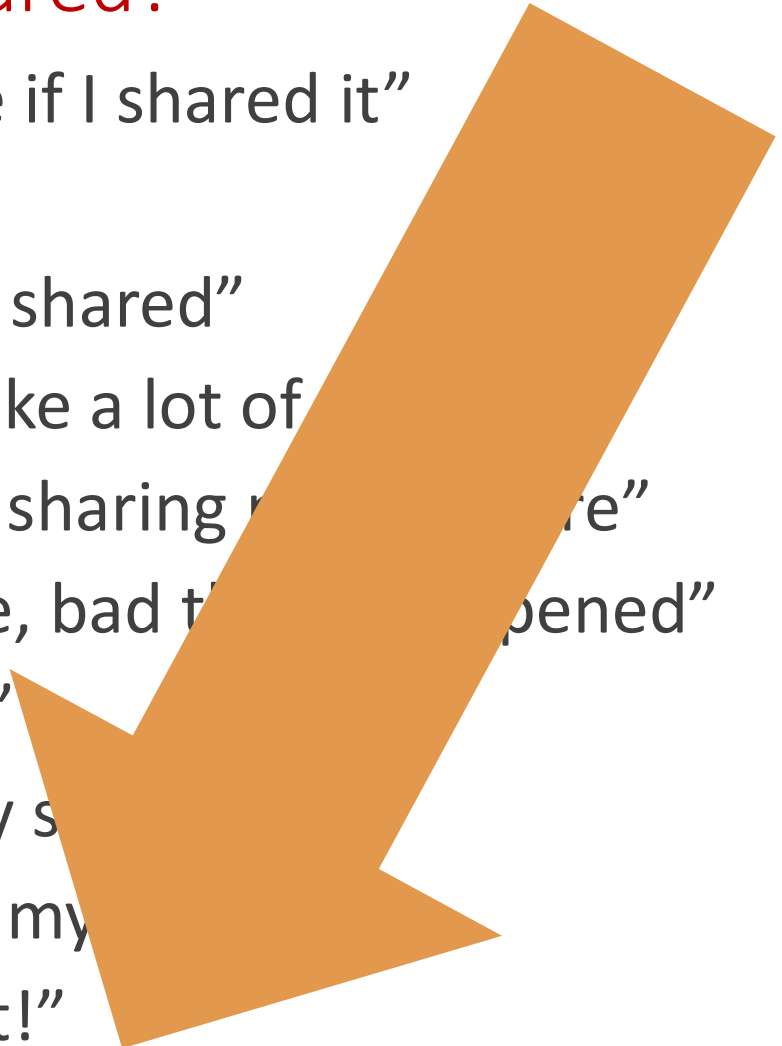
“Dark Software”



- 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?

- “No one would use my code if I shared it”
 - “My code is really bad”
 - “My code is not ready to be shared”
 - “Sharing my software will take a lot of time”
 - “I won’t get anything out of sharing my software”
 - “I’ve shared software before, bad things happened”
 - “I work for the government”
 - “I want to commercialize my software”
 - “I don’t want anyone to sell my software”
 - “I don’t know where to start!”
- 

Software Repository vs Software Registry

- Software repository

- Code resides there
- Support software evolution
- Support groups of developers of open source software

- Software registry

- Capture metadata
- Useful structured information about the code



OntoSoft



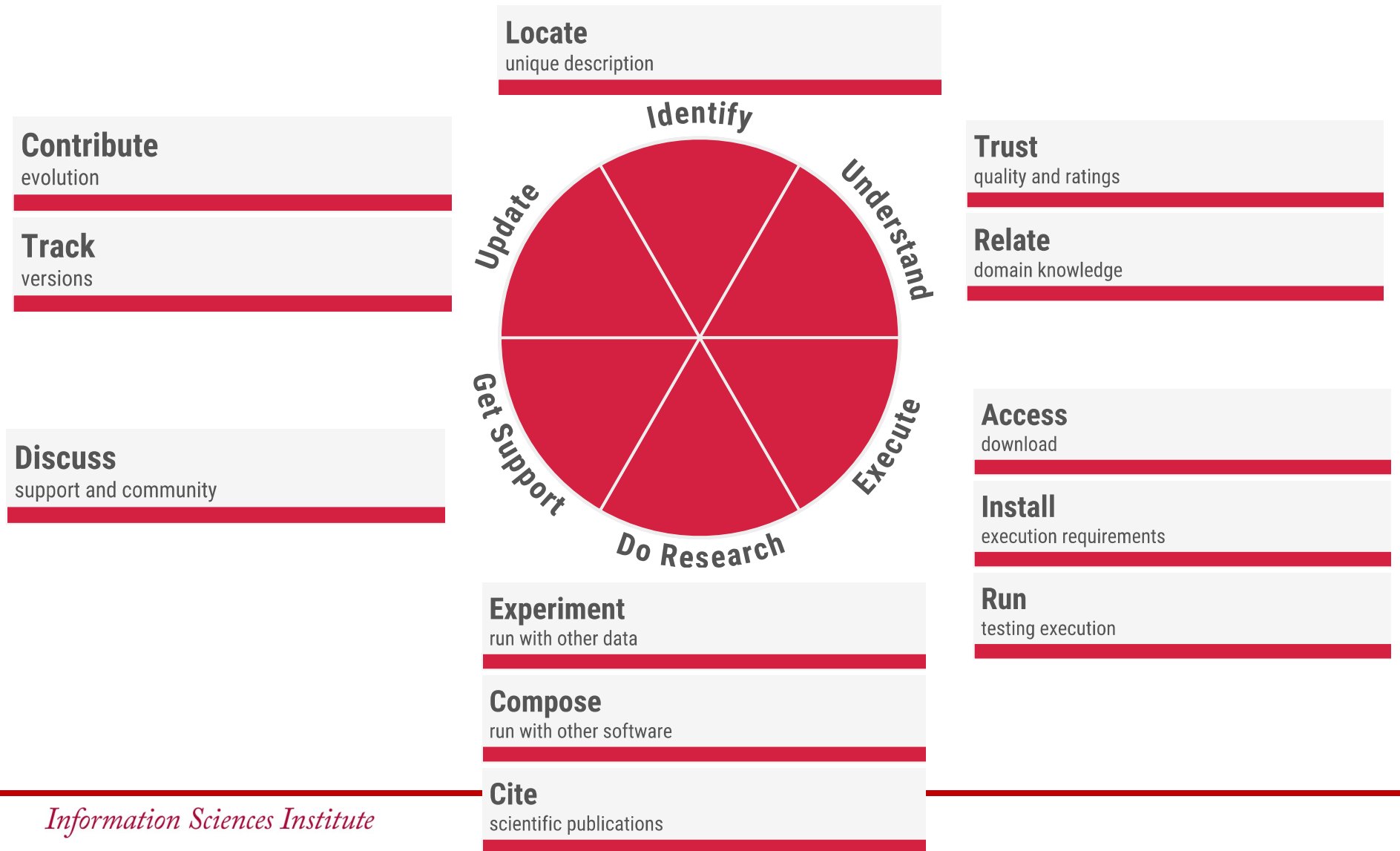
<http://ontosoft.org>

<http://imcr.ontosoft.org>

Registry for software

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

OntoSoft Metadata Categories



Describing Scientific Software in OntoSoft

The screenshot displays the OntoSoft portal interface for describing scientific software. The top navigation bar includes 'OntoSoft', 'Software', and 'Community' links, along with a user profile 'admin'. The main content area is titled '3DDY > Edit > Execute > INSTALL'. A circular progress indicator on the left shows the 'Identify' phase completed, with other phases like 'Understand', 'Execute', 'Do Research', 'Get Support', and 'Update' remaining. A 'SAVE' button is visible. The 'INSTALL' section is divided into 'Important' and 'Optional' categories. The 'Important' category includes questions like 'Is there any on-line documentation about the software?', 'What language(s) is the software written in?', 'What Operating Systems can the software run on?', and 'How can one install the software?'. The 'Optional' category includes 'What other software does the software require to be installed?'. Callouts provide additional context: 'Metadata properties organized into categories that make sense to scientists' points to the 'Important' category; 'Metadata properties collected through simple questions' points to the 'Optional' category; 'Automatic import of metadata from other repositories' points to the 'GDAL framework package 1.11' entry; and 'Indicators of metadata completeness' points to the circular progress indicator.

OntoSoft Software Community admin

3DDY > Edit > Execute > **INSTALL**

Identify Understand Execute Do Research Get Support Update

Access download

Install execution / requirements

Run testing execution

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

Automatic import of metadata from other repositories

Indicators of metadata completeness

Access control

Set Permissions for 3DDY

User

Permission

☐ Owner

Browse Permissions

Username	Write	Owner
No Permissions found..		

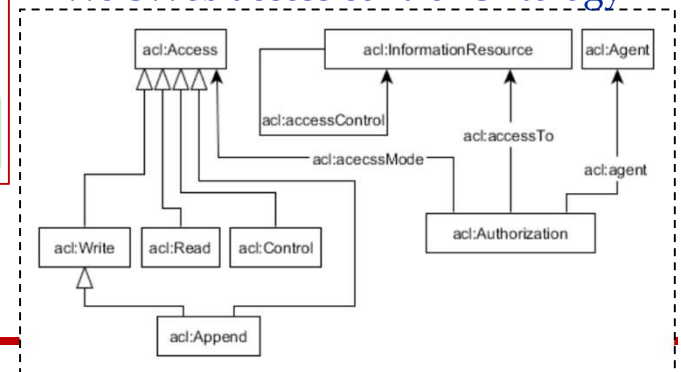
1-1 of 0

CANCEL **SUBMIT**

Setting permissions for editing 3DDY metadata

Users and permissions for the 3DDY software component

W3CWeb access control Ontology



Finding Software with OntoSoft

The screenshot displays the OntoSoft Software Repository interface. At the top, there is a navigation bar with links for 'OntoSoft', 'Software', 'Community', and 'Training'. Below this, the main heading is 'Software Repository' with the subtitle 'Describe your software so others can find and use it'. A button labeled 'PUBLISH YOUR SOFTWARE' is visible. The central part of the interface is divided into two main sections: 'Software List' and 'Filter Software List'. The 'Software List' section contains a table with the following entries:

Name	Actions
DrEICH algorithm	EDIT
PIHM	EDIT
PIHMgis	EDIT
TauDEM	EDIT
WBMsed	EDIT

The 'Filter Software List' section on the right includes a search bar and several filter buttons: 'Author', 'Keywords: Hydrological model OR Hydrology', 'Language: C++', and 'License: GNU General Public License v2.0'. At the bottom of the filters, there is a dropdown menu showing 'GNU General Public Lice' with an upward arrow and a close button.

Currently >600 entries, many imported from CSDMS, C4P, ...

Comparing Alternatives with OntoSoft

<div> <div>OntoSoft</div> <div>Software</div> <div>Community</div> <div>Training</div> </div> <div> <h2>Compare Software</h2> <p>DrEICH algorithm, PIHM, PIHMgis, TauDEM, WBMsed</p> </div>				
PIHM	PIHMgis	DrEICH	TauDEM	WBMsed
What are domain specific keywords for this software ? (eg: hydrology, climate)				
Geomorphology, Hydrological, Bedrock channel ero-	Basins, Continental	Basins, GIS	Hydrologically corrected DEM, Watershed	Sediment flux, Global model, Hydrological model
What Operating Systems can the software run on ?				
Unix Linux	Unix Windows Linux Mac OS	Unix Windows Linux Mac OS	Unix Windows Linux Mac OS	Unix Linux
Is there any test data available for the software ?				
Test Data Location: http://onlinelibrary.wiley.com/doi/10.1002/2013WR015167/full Test Data Description: Two test DEMs are included in the repository,	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://csdms.colorado.edu/wiki/Model:TauDEM#Testing Test Data Description: The Logan River DEM is a small test dataset useful	Test Data Location: http://csdms.colorado.edu/wiki/Model:WBMsed#Testing Test Data Description: Extensive input dataset is available on the CSDMS

Select software and features, get a comparison table

Publishing Software Metadata with OntoSoft

The screenshot shows the OntoSoft web interface for publishing software metadata. The header includes the OntoSoft logo and a menu icon. The main content area is titled "PIHM" with the author "[Christopher Duffy]". Below the title, there are three tabs: "HTML", "RDF/XML", and "JSON". The "HTML" tab is selected and highlighted with a red circle. To the right of the tabs is a "RATE" button. Below the tabs, the "Identify" section is visible, followed by the "Locate - Unique description" section. The "Locate" section contains two input fields: "What is the software called ?" and "What is a short description for this software ?". The first input field has a dropdown menu with "PIHM" selected. The second input field has a text area with the description: "PIHM is a multiprocess, multi-scale hydrologic model where the major coupled using the semi-discrete finite volume method. PIHM is a physio".

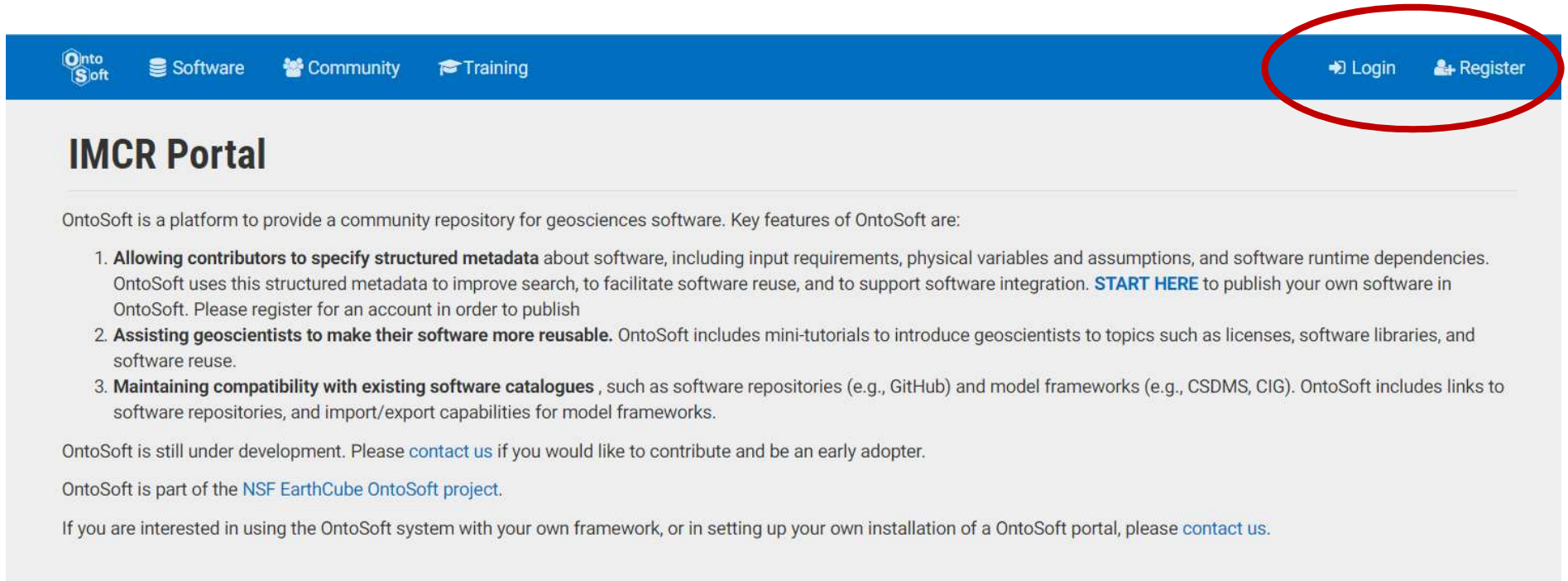
Publish metadata as HTML from OntoSoft and add pointer from software repository

OntoSoft Tutorial

How to use OntoSoft in a few steps

Registering in OntoSoft

- Go to <http://imcr.ontosoft.org>



The screenshot shows the OntoSoft IMCR Portal website. The top navigation bar is blue and contains links for OntoSoft, Software, Community, and Training. On the right side of the bar, 'Login' and 'Register' buttons are circled in red. The main content area is titled 'IMCR Portal' and describes the platform's features and goals.

IMCR Portal

OntoSoft is a platform to provide a community repository for geosciences software. Key features of OntoSoft are:

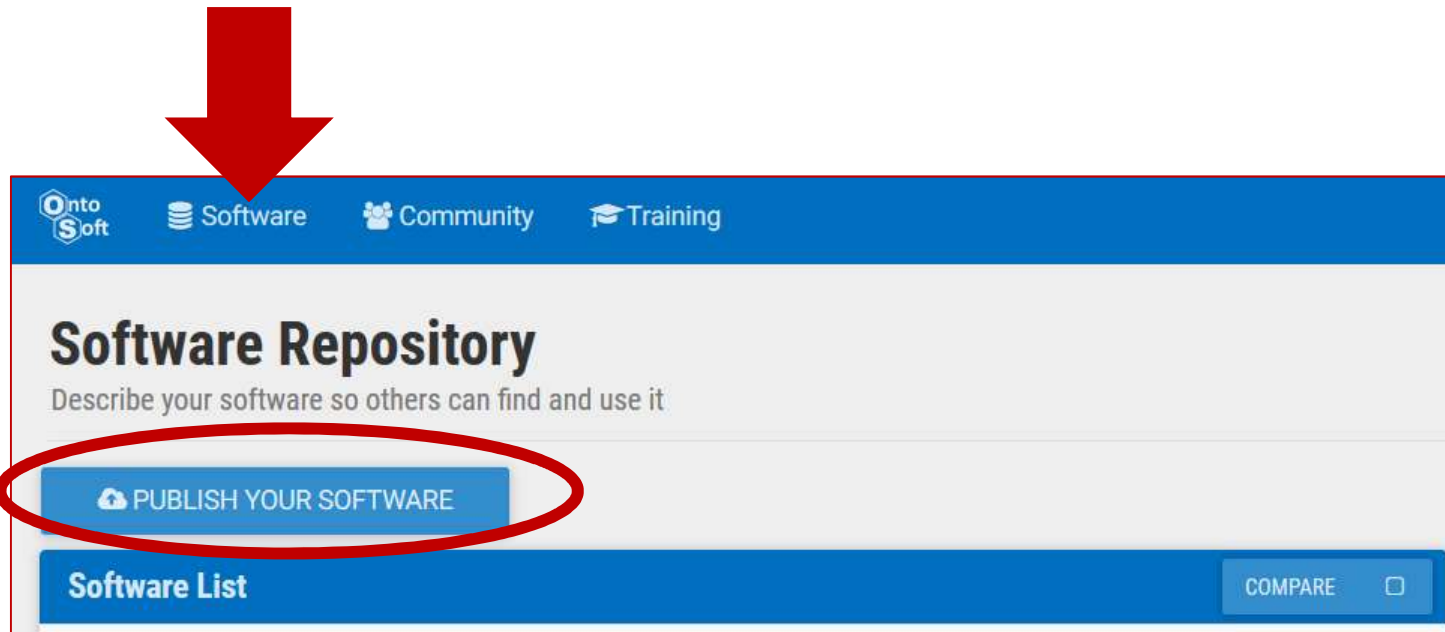
1. **Allowing contributors to specify structured metadata** about software, including input requirements, physical variables and assumptions, and software runtime dependencies. OntoSoft uses this structured metadata to improve search, to facilitate software reuse, and to support software integration. [START HERE](#) to publish your own software in OntoSoft. Please register for an account in order to publish
2. **Assisting geoscientists to make their software more reusable.** OntoSoft includes mini-tutorials to introduce geoscientists to topics such as licenses, software libraries, and software reuse.
3. **Maintaining compatibility with existing software catalogues**, such as software repositories (e.g., GitHub) and model frameworks (e.g., CSDMS, CIG). OntoSoft includes links to software repositories, and import/export capabilities for model frameworks.

OntoSoft is still under development. Please [contact us](#) if you would like to contribute and be an early adopter.

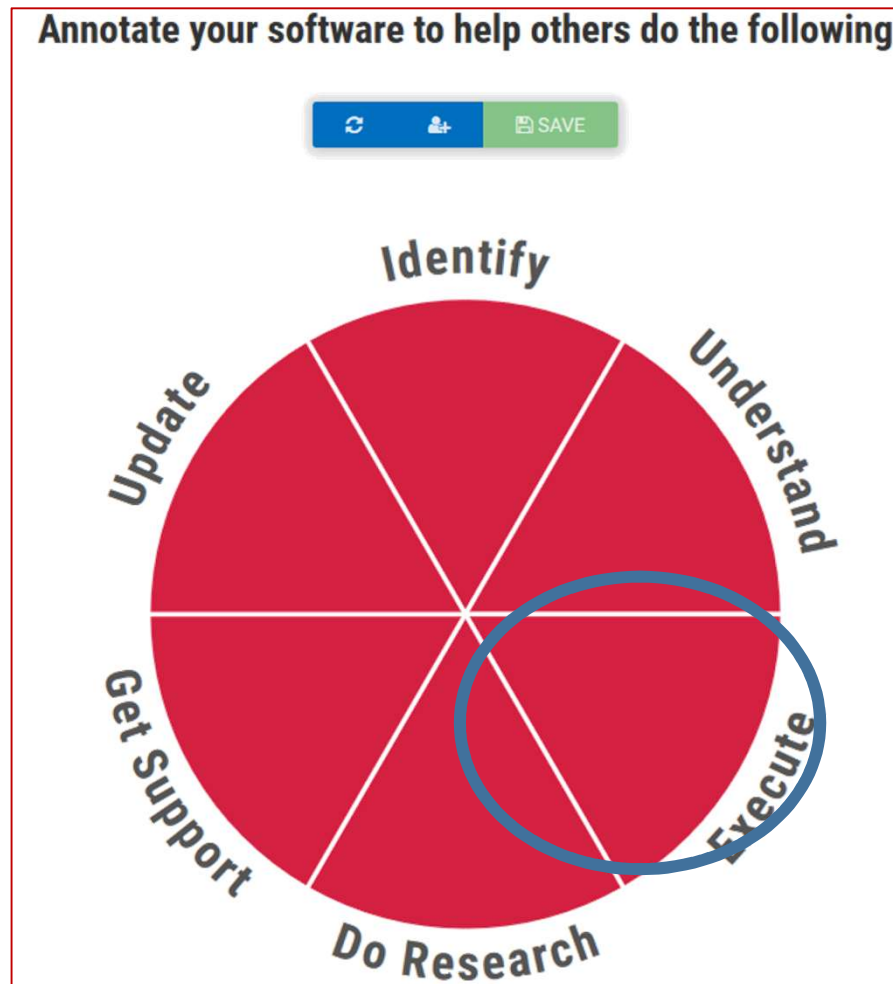
OntoSoft is part of the [NSF EarthCube OntoSoft project](#).

If you are interested in using the OntoSoft system with your own framework, or in setting up your own installation of a OntoSoft portal, please [contact us](#).

Adding a software entry



Adding a software entry (2)



- Fill in the different metadata fields by clicking on the different sections.
- The red color indicates the level of metadata completion of a category

Importing metadata from GitHub

The screenshot shows the WIDOCO 'ACCESS' interface. At the top, a breadcrumb trail reads 'WIDOCO » Edit » Execute » ACCESS'. Below this, a blue banner states 'Found some information from Github'. The main area features a circular diagram with six segments: 'Identify' (green), 'Understand' (pink), 'Execute' (red), 'Do Research' (pink), 'Get Support' (pink), and 'Update' (pink). To the right of the diagram is a form with two tabs: 'Important' and 'Optional'. The 'Important' tab is active, showing two questions: 'What is the URL for the code ?' with the answer 'https://github.com/dgarijo/Widoco/' and 'Last edited by admin at 2018-06-09 21:01', and 'What license is the code released under ?' with the answer 'License (License)'. A blue button labeled 'GET GITHUB METADATA' is positioned between the two questions. At the bottom left, there are three sections: 'Access' (download), 'Install' (execution requirements), and 'Run' (testing execution).

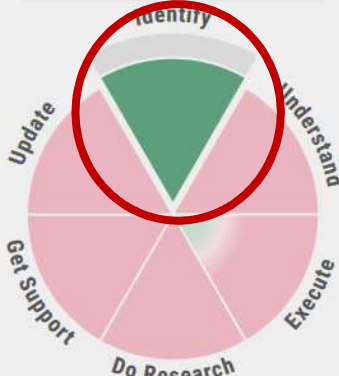
- Select “Execute” and then “Access”
- Insert your GitHub project URL on the first field
- Click on “Get GitHub Metadata”

Describing a software entry

WIDOCO » Edit » Identify » **LOCATE**

Found some information from Github

Refresh Add SAVE



Locate
unique description

Important **Optional**

What is the software called ? ←

WIDOCO

What is a short description for this software ?

A wizard for documenting ontologies automatically

What are general categories (keywords, labels) for this software ? →

Documentation

Is there a project website for the software ?

<https://w3id.org/widoco/>

1. Click on a category you want to describe (e.g., identify)
2. Answer the questions
3. Some questions allow for adding multiple responses (e.g., keywords)

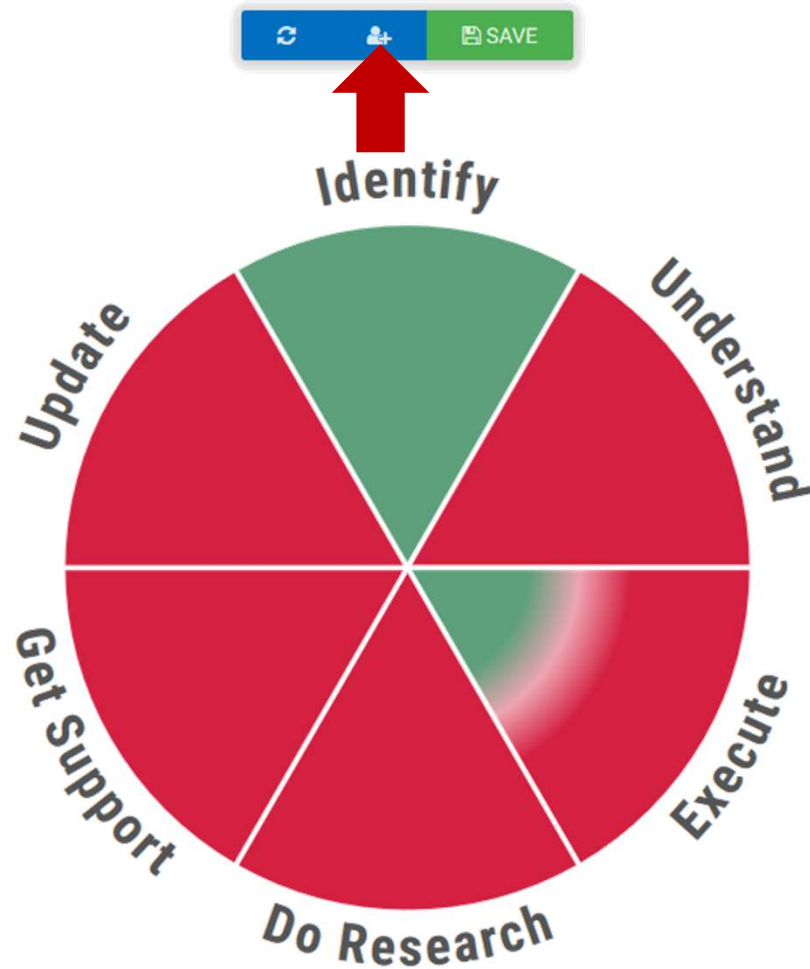
Describing a software entry (2)

The screenshot shows the WIDOCO software entry interface. At the top, there is a navigation bar with 'WIDOCO', 'Edit', 'Execute', and 'ACCESS'. Below this, a blue banner reads 'Found some information from Github'. The main area is divided into two tabs: 'Important' and 'Optional'. The 'Important' tab is active, showing a question 'What is the URL for the code ?' with the answer 'https://github.com/dgarijo/Widoco/'. Below this, there is a button 'GET GITHUB METADATA'. The 'Optional' tab is also visible, showing a question 'What license is the code released under ?' with a dropdown menu of licenses: 'Apache License 2.0', 'Common Development and Distribution License', and 'GNU Affero General Public License v3.0'. On the left side, there is a circular diagram with six segments: 'Identify', 'Understand', 'Execute', 'Do Research', 'Get Support', and 'Update'. Below the diagram, there are three sections: 'Access' (download), 'Install' (execution requirements), and 'Run' (testing execution).

Some fields allow autocompleting metadata from existing information (e.g., licenses)
Just start writing to show if there are existing licenses that overlap with yours

Allowing others to change your metadata

Annotate your software to help others do the following



Allowing others to change your metadata (2)

Set Permissions for WIDOCO ×

Usercgries

PermissionWrite

☐ Owner

Default Permission: Read

Browse Permissions

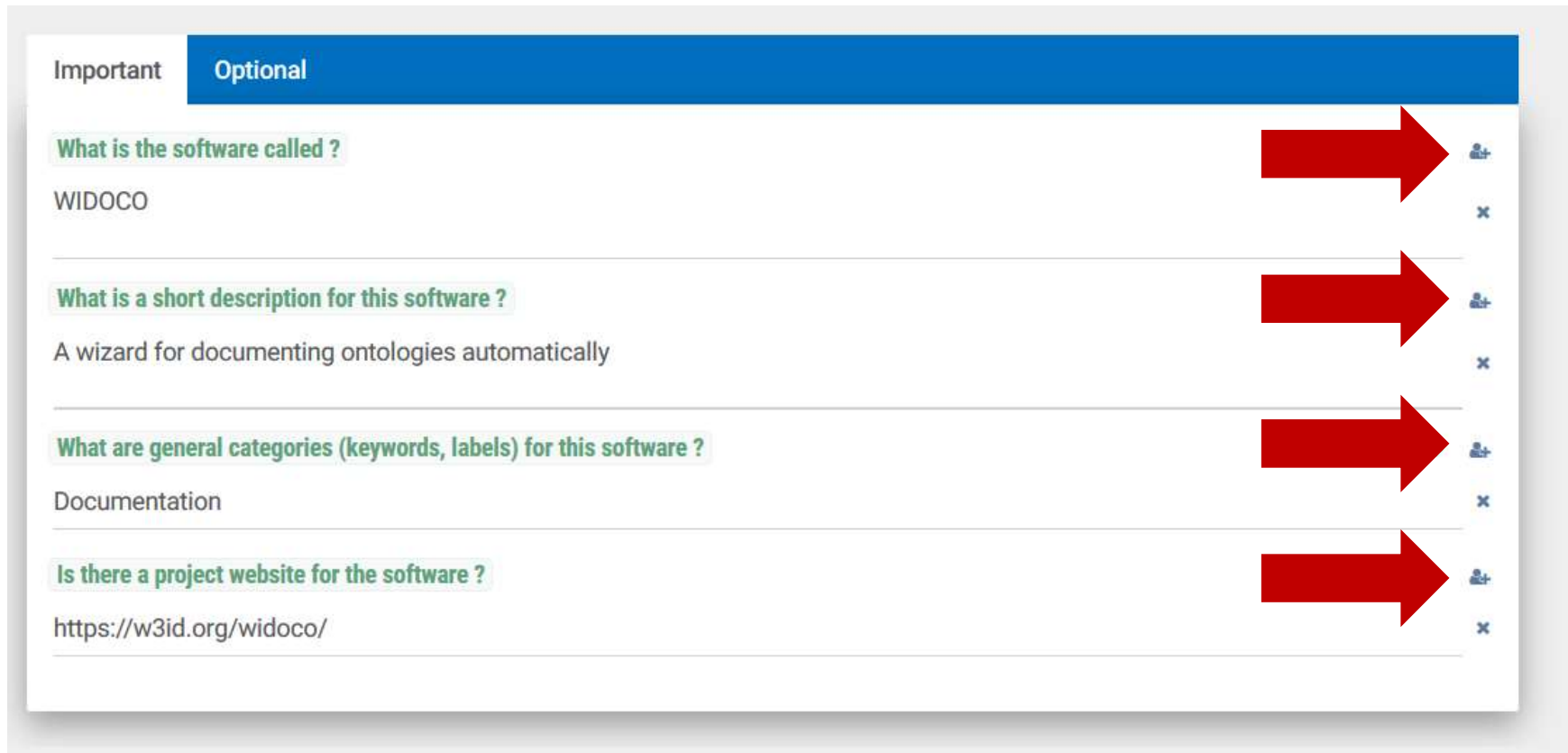
▲ Username	Write	Owner
admin	✓	✓

⏮ ⏪ 1-1 of 1 ⏩ ⏭

CANCEL

SUBMIT

Allowing others to change your metadata (3)



The screenshot shows a metadata form with two tabs: 'Important' and 'Optional'. The 'Optional' tab is selected. There are four input fields, each with a red arrow pointing to a user icon on the right, indicating that these fields can be edited by other users.

Field Label	Value	Editable by others
What is the software called ?	WIDOCO	Yes
What is a short description for this software ?	A wizard for documenting ontologies automatically	Yes
What are general categories (keywords, labels) for this software ?	Documentation	Yes
Is there a project website for the software ?	https://w3id.org/widoco/	Yes

You can set individual properties to be changed by another user

Downloading your software metadata



WIDOCO [No author listed]

IDENTIFY

Locate - Unique description

What is the software called ?

- WIDOCO

What is a short description for this software ?

- A wizard for documenting ontologies automatically

What are general categories (keywords, labels) for this software ?

- Documentation

Is there a project website for the software ?

- <https://w3id.org/widoco/>

EXECUTE

Access - Download

What is the URL for the code ?

- <https://github.com/dgarijo/Widoco/>

1. Click on the serialization you want to preserve
2. Save it to your disk!

Finding Software with OntoSoft

The screenshot shows the OntoSoft Software Repository website. The header includes the OntoSoft logo and navigation links for Software, Community, and Training. The main heading is 'Software Repository' with the tagline 'Describe your software so others can find and use it'. A prominent button says 'PUBLISH YOUR SOFTWARE'. Below this, there's a 'Software List' section with a 'COMPARE' button. The list contains five entries: DrEICH algorithm, PIHM, PIHMgis, TauDEM, and WBMsed, each with an 'EDIT' button. To the right is a 'Filter Software List' sidebar with a search bar and several filter buttons: 'Author', 'Keywords: Hydrological model OR Hydrology', 'Language: C++', and 'License: GNU General Public License v2.0'. The bottom of the sidebar shows a partial filter for 'GNU General Public License'.

OntoSoft Software Community Training

Software Repository

Describe your software so others can find and use it

PUBLISH YOUR SOFTWARE

Software List

COMPARE

Name	Actions
DrEICH algorithm	EDIT
PIHM	EDIT
PIHMgis	EDIT
TauDEM	EDIT
WBMsed	EDIT

Filter Software List

Search

Author

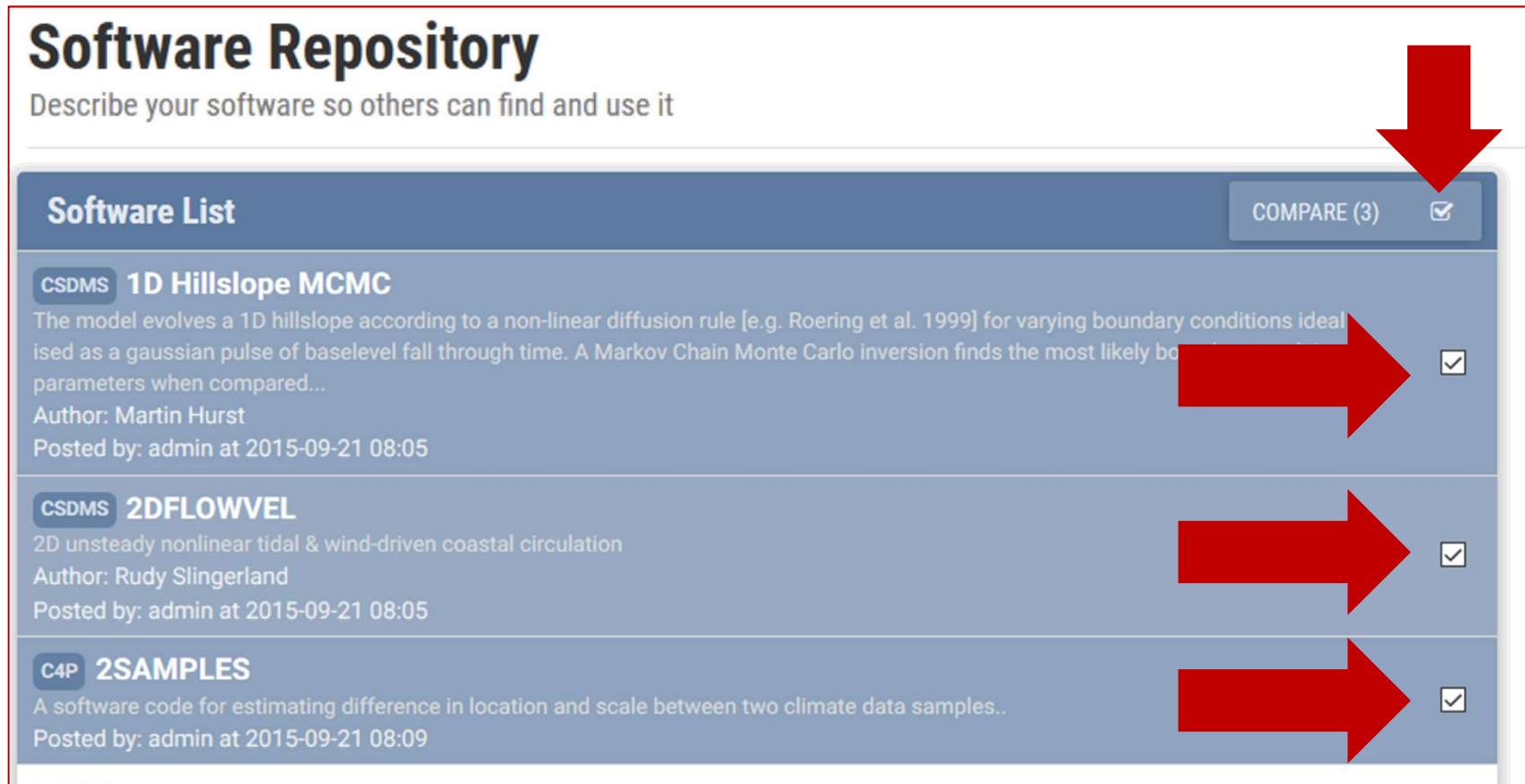
Keywords: Hydrological model OR Hydrology

Language: C++

License: GNU General Public License v2.0

GNU General Public License

Comparing Software



Software Repository
Describe your software so others can find and use it

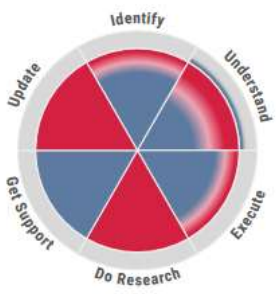
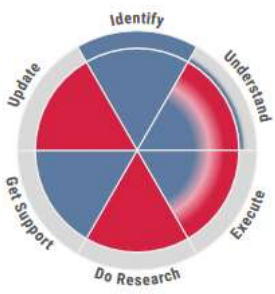

Software List COMPARE (3) ☒

CSDMS 1D Hillslope MCMC The model evolves a 1D hillslope according to a non-linear diffusion rule [e.g. Roering et al. 1999] for varying boundary conditions idealised as a gaussian pulse of baselevel fall through time. A Markov Chain Monte Carlo inversion finds the most likely parameters when compared... Author: Martin Hurst Posted by: admin at 2015-09-21 08:05	<input checked="" type="checkbox"/>
CSDMS 2DFLOWVEL 2D unsteady nonlinear tidal & wind-driven coastal circulation Author: Rudy Slingerland Posted by: admin at 2015-09-21 08:05	<input checked="" type="checkbox"/>
C4P 2SAMPLES A software code for estimating difference in location and scale between two climate data samples.. Posted by: admin at 2015-09-21 08:09	<input checked="" type="checkbox"/>

The screenshot shows a web interface for a software repository. At the top, there's a header 'Software Repository' with a subtitle 'Describe your software so others can find and use it'. Below this is a 'Software List' section. On the right of this section, there is a 'COMPARE (3)' button with a checked checkbox. A large red arrow points down to this button. The list contains three entries, each with a title, description, author, and post date. To the right of each entry is a checkbox, and a large red arrow points to each of these checkboxes. The first entry is 'CSDMS 1D Hillslope MCMC', the second is 'CSDMS 2DFLOWVEL', and the third is 'C4P 2SAMPLES'. All three checkboxes are checked.

1. Click on “compare” box to enable the comparison mode
2. Select entries (click on their boxes)
3. Click on “compare” button

Comparing Software (2)

Compare Software		
1D Hillslope MCMC, 2DFLOWVEL, 2SAMPLES		
1D Hillslope MCMC	2DFLOWVEL	2SAMPLES
		
What is the software called ?		
1D Hillslope MCMC	2DFLOWVEL	2SAMPLES
What is a short description for this software ?		
<p>The model evolves a 1D hillslope according to a non-linear diffusion rule [e.g. Roering et al. 1999] for varying boundary conditions idealised as a gaussian pulse of baselevel fall through time. A Markov Chain Monte Carlo inversion finds the most likely boundary condition parameters when compared to a time series of field data on hillslope morphology from the Dragon's Back Pressure Ridge, Carrizo Plain, CA, USA [see Hilley and Arrowsmith, 2008].</p> <p>Initial metadata was retrieved from http://csdms.colorado.edu/wiki/Model:1D_Hillslope_MCMC</p>	<p>2D unsteady nonlinear tidal & wind-driven coastal circulation</p> <p>Initial metadata was retrieved from http://csdms.colorado.edu/wiki/Model:2DFLOWVEL</p>	<p>A software code for estimating difference in location and scale between two climate data samples..</p> <p>Initial metadata was retrieved from: https://www.ncdc.noaa.gov/cdo/f?p=517:20:0:::PROXYDATASETLIST:59</p>
What are general categories (keywords, labels) for this software ?		
Terrestrial Landscape evolution Hillslope Model	Coastal Flow dynamics	Climatology and Atmospheric Paleoclimatology Climatology Paleoenvironment
Is there a project website for the software ?		



ICER-1440323
ICER-1343800

Acknowledgements

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

- 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