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
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.

THIS IS GOING TO BE PAINFUL, ISN'T IT?

JUST A SCRATCH.
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

Comparison of dissimilar protein structures:

- Get Significant Results
- FATCAT URLChecker
- FATCAT
- Remove Significant Pairs

Docking:

- CreateClip Files
- CreateIdeal Ligands
- IdealLigand Checker
- Autodock Vina

Minimal
Novice
Author
Software Today

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

- 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?

- “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!”
Contributions: OntoSoft

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

- **Contribute**: evolution
- **Track**: versions
- **Discuss**: support and community

**Locate**
unique description

**Identify**

**Update**

**Get Support**

**Do Research**

**Understand**

**Execute**

**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/software
Describing Scientific Software in OntoSoft

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

Metadata can be exported in several formats (HTML, RDF, JSON)

Metadata for 3DDY Software

Set permissions for 3DDY

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

http://www.ontosoft.org/portal
Access control

Setting permissions for editing 3DDY metadata

Users and permissions for the 3DDY software component

W3CWeb access control Ontology

http://www.ontosoft.org/portal
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

Critical Zone Observatory

EarthCube RCNs

Oomics

Publication

CSDMS CIG ESMF

Learning

EarthCube Building Blocks FES/ESIP

Software Carpentry

Early Career Advisory Board

UK Software Institute

Rec o m m end ers

Recommender system 

Interoperability 
Publication 
Community 
Learning 
Structured metadata 
Interactive advice 
Best practices 
Multimedia lessons

Omics

EarthCube

SEN C4P EC3

Code meta initiative

Yolanda Gil, Daniel Garijo, Saurabh Mishra, Varun Ratnakar

eScience 2016

USC Information Sciences Institute
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

Do you want to use Ontosoft? Let us know!

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


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