

# The Scientific Paper of the Future

<http://www.scientificpaperofthefuture.org>

## OntoSoft Training

February 2017

[ontosoft@gmail.com](mailto:ontosoft@gmail.com)

<http://dx.doi.org/10.5281/zenodo.159206>



**Onto**  
**Soft**

NSF ER-1440323  
ER-1343800

CC-BY  
Attribution



 EarthCube

# Instructors Today

**Daniel  
Garijo**



**Yolanda Gil**



**Information Sciences Institute  
University of Southern California**



# Software Stewardship for the Geosciences



## Community

- Recommender system
- Interoperability



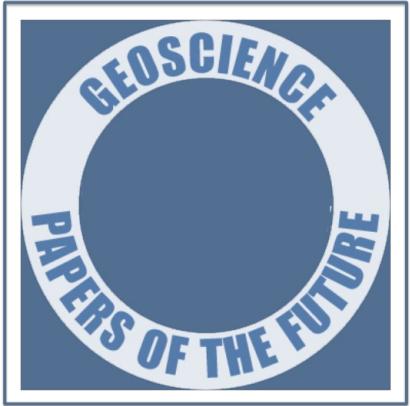
## Publication

- Structured metadata
- Interactive advice



## Learning

- Best practices
- Multimedia lessons



# The Geoscience Papers of the Future (GPF) Initiative

<http://www.scientificpaperofthefuture.org/gpf>

1. A Special Issue of a journal in all geoscience areas that includes only geoscience papers of the future



2. Training sessions for geoscientists to learn best practices in software and data sharing, provenance documentation, and scholarly publication

# GPF Pioneer Authors



**Cedric David**, NASA/JPL  
Hydrology modeling



**Ibrahim Demir**, U. of Iowa  
Hydrology sensor networks



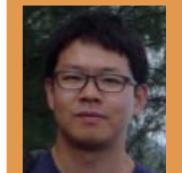
**R. W. Fulweiler**, Boston U.  
Biogeochemistry in marine ecology



**J. Goodall/B. Essawy**, U.  
Virginia, Hydrology/visualization



**Leif Karlstrom**, U. Oregon  
Volcanic vent clustering



**Kyo Lee**, NASA/JPL  
Regional climate modeling



**Heith Mills**, U. Houston  
Geochemistry, marine biology



**Ji-Hyun Oh**, USC  
Tropical meteorology



**Suzanne Pierce**, UT Austin  
Hydrogeology for decision supp



**Allen Pope**, U. Colorado  
Glaciology



**Mimi Tzeng**, Dauphin Island  
Sea Lab, Ocean fisheries



**Sandra Villamizar**, UC Merced  
River ecohydrology



**Xuan Yu**, U. Delaware  
Hydrologic modeling

# Why Learn to Write a Scientific Paper of the Future

1. **Get credit** for all your research products
  - ★ Citations for software, data, samples, ...
2. **Increase citations** of your papers
3. Write impressive **Data Management Plans**
4. **Extend your CV** with data and software sections
5. **Reproduce** your work from years ago
6. Comply with new **funder and journal requirements**



# Training Goals

## What Training Covers

- ★ **Best practices**
  - ★ Many are still being developed by the community
- ★ **Major concepts and goals**, regardless of the platform, research area, or target journal
- ★ **Mindful of effort**
  - ★ How to implement best practices with simplest approach

## What is Not Covered

- ★ Metadata standards specific to particular research areas
- ★ Improving software development skills
- ★ Details of using code sharing sites



# Scientific Paper of the Future Training

## Part I

1. Motivation and overview: open science, reproducible publications, and digital scholarship
2. Making data accessible
3. Making software accessible
4. Documenting software with metadata

## Part II

5. Documenting provenance and methods
6. Improving author citation profile and researcher impact
7. Summary of author checklist



**CODATA**



# The Scientific Paper of the Future: Motivation and Overview

## OntoSoft Training

### Part 1

<http://dx.doi.org/10.5281/zenodo.15920>



<http://www.scientificpaperofthefuture.org>

CC-BY  
Attribution



# Scientists Are Changing

Open data



Open source



Open access



Open publications



# Publishers Are Changing



## Illuminating the black box

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 information, please read [Reporting Life Sciences Research](#).

The logo for the journal 'nature' is shown in its signature red serif font.

## Availability of Software



PLOS supports the development of open source software and believes that, for submissions appropriate open source standards will ensure that the submission conforms to (1) our requirements that another researcher can reproduce the experiments described, (2) our aim to promote openness, and (3) that PLOS journals can be built upon by future researchers. Therefore, if new software or a new application is developed during the course of a study, and it is believed that the software conforms to the [Open Source Definition](#), have deposited the following three items as Supporting Information:

- **The associated source code of the software described by the paper.** This should be licensed under a suitable license such as BSD, LGPL, or MIT (see <http://www.opensource.org/licenses/>). The use of commercial software such as Mathematica and MATLAB does not preclude a paper from being considered open source, but it is preferred.
- **Documentation for running and installing the software.** For end-user applications, this may include instructions for installing the software and any prerequisites; for software libraries, instructions for using the application program interface.
- **A test dataset with associated control parameter settings.** Where feasible, results should be presented in a way that allows them to be reproduced. Test data should not have any dependencies — for example, a database dump.

Acceptable archives should provide a public repository of the described software. The code should be available without requiring users to create accounts, log in, or otherwise register personal details. The repository should contain more than 1,000 projects. Examples of such archives are: [SourceForge](#), [Bioinformatics.Org](#), [GitHub](#), [Githannah](#), [GitHub](#) and the [Codehaus](#). Authors should provide a direct link to the deposited software.

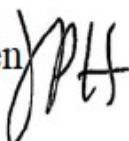
# Funders Are Changing

EXECUTIVE OFFICE OF THE PRESIDENT  
OFFICE OF SCIENCE AND TECHNOLOGY POLICY  
WASHINGTON, D.C. 20502

February 22, 2013

## MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: John P. Holdren  
Director



SUBJECT: Increasing Access to the Results of Federally Funded Scientific Research

### **1. Policy Principles**

The Administration is committed to ensuring that, to the greatest extent and with the fewest constraints possible and consistent with law and the objectives set out below, the direct results of federally funded scientific research are made available to and useful for the public, industry, and the scientific community. Such results include peer-reviewed publications and digital data.

an approach for optimizing search, archival, and dissemination features that encourages innovation in accessibility and interoperability, while ensuring long-term stewardship of the results of federally funded research;

# Modern Scientific Articles

## Traditional Published Articles

### **Text:**

Narrative of method,  
the data is in tables, figures/plots,  
the software used is mentioned



## Modern Published Articles

### **Text:**

Narrative of method,  
the data is in tables, figures/plots,  
the software used is mentioned

### **Data:**

Supplementary materials,  
pointers to data repositories

# Data Papers & Data Repositories

## ★ Data paper

Ecological Research  
July 2013, Volume 28, Issue 4, p 541

Date: 10 May 2013

### Monitoring records of plant species in the Hakone region of Fuji-Hakone-Izu National Park, Japan, 2001–2010

Takeshi Osawa



#### Abstract

The monitoring of species occurrences is a crucial aspect of biodiversity conservation, and regional volunteerism can serve as a powerful tool in such endeavors. The Fuji-Hakone-Izu National Park in the Hakone region of Kanagawa Prefecture, Japan, boasts a volunteer association of approximately 100 members. These volunteers have monitored plant species occurrences from 2001 to the present along several hiking trails in the region. In this paper, I present the annual observation records of plant occurrences in Hakone from 2001 to 2010. This data set includes 1,071 species of plants from 151 families. Scientific names follow the Y List, and this data set includes several threatened plant species. Data files are formatted based on the Darwin Core and Darwin Core Archives, which are defined by the Biodiversity Information Standards (BIS) or Biodiversity Information Standards Taxonomic Databases Working Group (TDWG). Data files filled on required and some additional item on Darwin Core. The data set can download from the author's personal Web site as of July 2012. These data will soon be published for the Global Biodiversity Information Facility (GBIF) through GBIF Japan. All users can then access the data from the GBIF portal site.

- The complete data set for this abstract published in the Data Paper section of the journal is available in electronic format in Ecological Research Data Paper Archives at [http://db.cger.nies.go.jp/JaLTER/ER\\_DataPapers/archives/2013/ERDP-2013-01](http://db.cger.nies.go.jp/JaLTER/ER_DataPapers/archives/2013/ERDP-2013-01).

## ★ Data published in a repository

The screenshot shows the LTER network interface. At the top, it says "The US Long Term Ecological Research Network". Below that, there's a search bar with "WDNR Yahara Lakes Fisheries: Fish Lengths and Weights 1987-1998 - Lathrop". To the left of the search bar is a logo for LTER. Below the search bar, it says "LTER Identifier: knb-lter-ndl.279.1". Under "Abstract:", it says "These data were collected by the Wisconsin Department of Natural Resources (WDNR) from 1987-1998. Most of these data (1987-1993) precede 1995, the year that the University of Wisconsin took over sampling of the Yahara Lakes. However, WDNR data collected from 1997-1998 (unrelated to LTER sampling) is also included. In 1987 a joint project by the WDNR and the University of Wisconsin-Madison, Center for Limnology (CFL) was initiated on Lake Mendota. The project involved biomanipulation o...".

#### LTER Identifier:

knb-lter-ndl.279.1

#### Abstract:

These data were collected by the Wisconsin Department of Natural Resources (WDNR) from 1987-1998. Most of these data (1987-1993) precede 1995, the year that the University of Wisconsin took over sampling of the Yahara Lakes. However, WDNR data collected from 1997-1998 (unrelated to LTER sampling) is also included. In 1987 a joint project by the WDNR and the University of Wisconsin-Madison, Center for Limnology (CFL) was initiated on Lake Mendota. The project involved biomanipulation o...

#### Owners/Creators:

Lathrop

#### Metadata:

Select [here](#) for full metadata

#### Data File(s):

- [wdnr\\_fyke\\_minifyke\\_seine\\_lengths\\_weights.csv](#)
- [wdnr\\_boomshock\\_lengths\\_weights.csv](#)
- [wdnr\\_gillnet\\_lengths\\_weights\\_93.csv](#)
- [wdnr\\_walleye\\_age\\_lengths\\_weights\\_87.csv](#)
- [wdnr\\_creek\\_survey\\_lengths\\_weights.csv](#)
- [wdnr\\_creek\\_survey\\_angler\\_counts.csv](#)

# “Dark Data”

## Shedding Light on the Dark Data in the Long Tail of Science

P. Bryan Heidorn

---

From: Library Trends

Volume 57, Number 2, Fall 2008

pp. 280-299 | 10.1353/lib.0.0036

---

### Abstract:

One of the primary outputs of the scientific enterprise is data, but many institutions such as libraries that are charged with preserving and disseminating scholarly output have largely ignored this form of documentation of scholarly activity. This paper focuses on a particularly troublesome class of data, termed *dark data*. "Dark data" is not carefully indexed and stored so it becomes nearly invisible to scientists and other potential users and therefore is more likely to remain underutilized and eventually lost. The article discusses how the concepts from long-tail economics can be used to understand potential solutions for better curation of this data. The paper describes why this data is critical to scientific progress, some of the properties of this data, as well as some social and technical barriers to proper management of this class of data. Many potentially useful institutional, social, and technical solutions are under development and are introduced in the last sections of the paper, but these solutions are largely unproven and require additional research and development.

# Modern Scientific Articles

## Traditional Published Articles



## Modern Published Articles

### **Text:**

Narrative of method,  
the data is in tables, figures/plots,  
the software used is mentioned

### **Text:**

Narrative of method,  
the data is in tables, figures/plots,  
the software used is mentioned

### **Data:**

Supplementary materials,  
pointers to data repositories

### **NOT published, loosely recorded:**

**Software:**  
scripted codes + manual steps +  
documentation in notes/emails

# Reproducibility

## Illuminating the black box

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 information, please read [Reporting Life Sciences Research](#).

nature

## COMPUTER SCIENCE

## Accessible Reproduc



## Methodology

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

By Paul Goldberg

Human lives

Biostatistics journals aren't usually the most recent issue of the Annals of Applied Statistics is an

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

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

The New York Times

NYTimes: Home - Site Index - Archive - Help

## Reliability

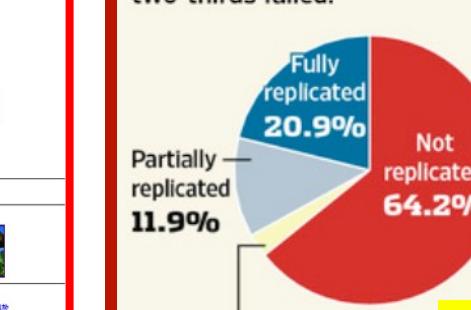
### Nobel Laureate Retracts Two Papers

The New York Times

## Retracted Scientific Studies: A Growing List

RETRACTED

## Scientific integrity



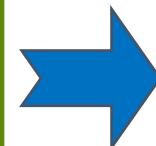
Source: Nature Reviews Drug Discov

## Trust



# Reproducible Articles

## Modern Published Articles



### **Text:**

Narrative of method,  
the data is in tables, figures/plots,  
the software used is mentioned

### **Data:**

Supplementary materials,  
pointers to data repositories

**NOT published,  
loosely recorded:**

### **Software:**

scripted codes + manual steps  
documentation in notes/emails

## Reproducible Publications

### **Text:**

Narrative of method,  
the data is in tables, figures/plots,  
the software used is mentioned

### **Data:**

Supplementary materials,  
pointers to data repositories

### **Software:**

Data preparation,  
data analysis, and visualization

### **Provenance and Workflow:**

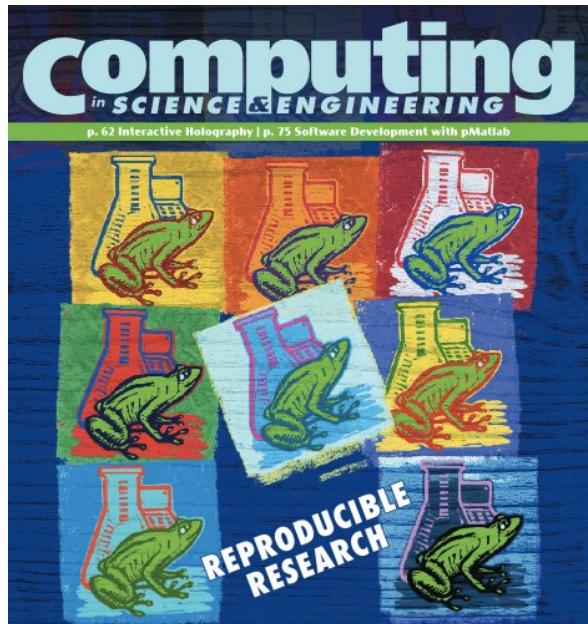
Workflow/scripts describing  
dataflow, codes, and parameters

# Reproducible Publications and Executable Papers



Sweave = R · L<sup>A</sup>T<sub>E</sub>X

IP[y]: Notebook



# Beyond Reproducible Publications

## Reproducible Publications

### Text:

Narrative of method,  
the data is in tables, figures/plots,  
the software used is mentioned

### Data:

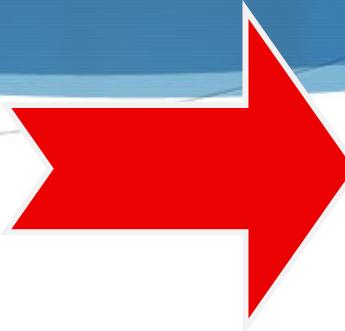
Supplementary materials,  
pointers to data repositories

### Software:

Data preparation,  
data analysis, and visualization

### Provenance and methods:

Workflow/scripts describing  
dataflow, codes, and parameters



Is this sufficient?

The Scientific  
Paper of the  
Future has further  
requirements

# Citations: Getting Credit

OPEN  ACCESS Freely available online

 PLOS one

## Sharing Detailed Research Data Is Associated with Increased Citation Rate

Heather A. Piwowar\*, Roger S. Day, Douglas B. Fridsma

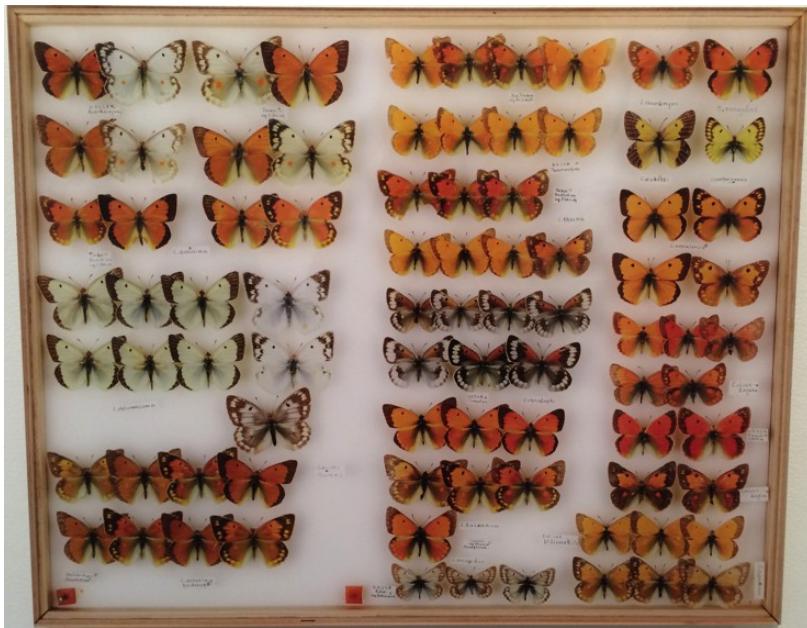
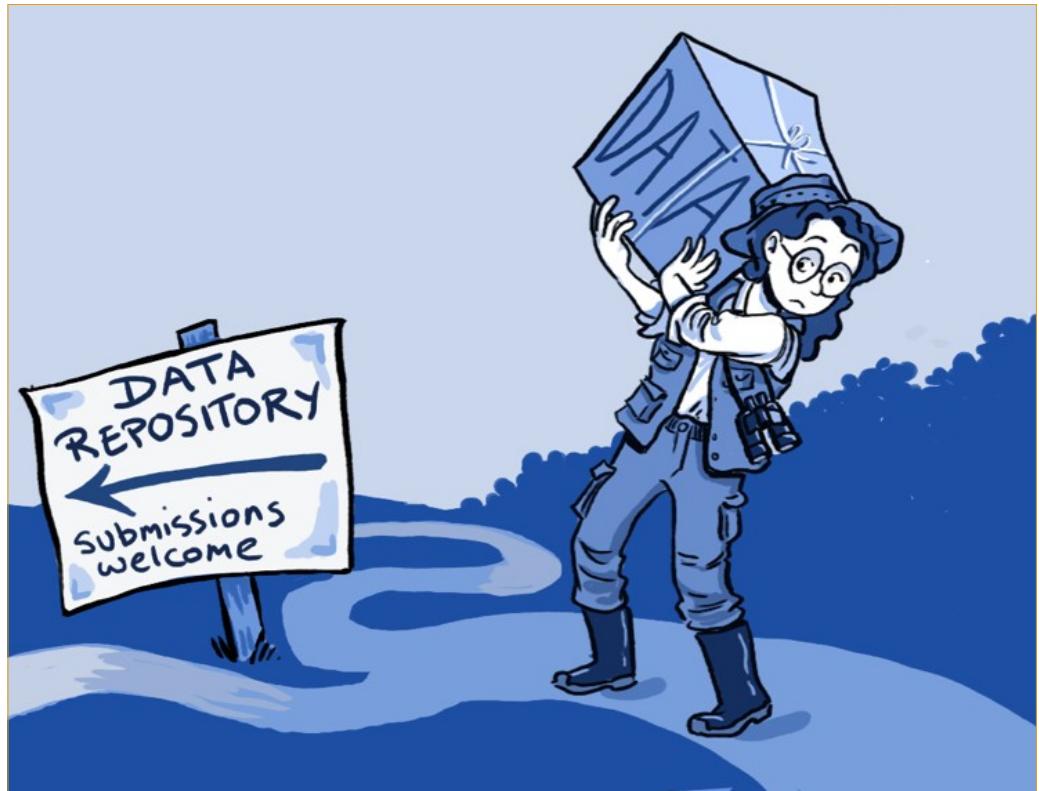
Department of Biomedical Informatics, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, United States of America

**Background.** Sharing research data provides benefit to the general scientific community, but the benefit is less obvious for the investigator who makes his or her data available. **Principal Findings.** We examined the citation history of 85 cancer microarray clinical trial publications with respect to the availability of their data. The 48% of trials with publicly available microarray data received 85% of the aggregate citations. Publicly available data was significantly ( $p=0.006$ ) associated with a 69% increase in citations, independently of journal impact factor, date of publication, and author country of origin using linear regression. **Significance.** This correlation between publicly available data and increased literature impact may further motivate investigators to share their detailed research data.

# Licenses for Data and Software: Encouraging Safe Reuse



# Discoverability through Shared Repositories and Metadata for Data and Software



# Scientific Paper of the Future

## Modern Paper

### **Text:**

Narrative of the method,  
some data is in tables,  
figures/plots, and the  
software used is mentioned

### **Data:**

Include data as  
supplementary materials  
and pointers to  
data repositories

## Reproducible Publication

### **Software:**

For data preparation, data  
analysis, and visualization

**Provenance and methods:**  
Workflow/scripts specifying  
dataflow, codes,  
configuration files,  
parameter settings, and  
runtime dependencies

## Open Science

### **Sharing:**

Deposit data and software  
(and provenance/workflow)  
in publicly shared repositories

### **Open licenses:**

Open source licenses for  
data and software  
(and provenance/workflow)

### **Metadata:**

Structured descriptions of the  
characteristics of data and software  
(and provenance/workflow)

## Digital Scholarship

### **Persistent identifiers:**

For data, software, and authors  
(and provenance/workflow)

### **Citations:**

Citations for data and software  
(and provenance/workflow)

# What is a Scientific Paper of the Future

- ★ **Data:** Available in a public repository, including documentation (metadata), a clear license specifying conditions of use, and citable using a unique and persistent identifier.
- ★ **Software:** Available in a public repository, with documentation (metadata), a license for reuse, and citable using a unique persistent identifier.
  - ★ Not only major software used, but also other ancillary software for data reformatting, data conversions, data filtering, and data visualization.
- ★ **Provenance:** Documented for all results by explicitly describing the series of computations and their outcome with a provenance record of the execution traces and a workflow sketch (or formal workflow)
  - ★ Possibly in a shared repository and with a unique and persistent identifier.

# Making Data

# Accessible

## OntoSoft Training

### Part 2

<http://dx.doi.org/10.5281/zenodo.15920>



<http://www.scientificpaperofthefuture.org>



**Onto**  
**Soft**

NSF ER-1440323  
ER-1343800

CC-BY  
Attribution



# Problems with Current Practice

- ★ Data is often not made available in publications
  - ★ Lack of reproducibility

*Nature Genetics* **41**, 149 - 155 (2009)  
Published online: 28 January 2008 | doi:10.1038/ng.295

## Repeatability of published microarray gene expression analyses

scientists. Here we evaluated the replication of data analyses in 18 articles on microarray-based gene expression profiling published in *Nature Genetics* in 2005–2006. One table or figure from each article was independently evaluated by two teams of analysts. We reproduced two analyses in principle and six partially or with some discrepancies; ten could not be reproduced. The main reason for failure to reproduce was data unavailability, and discrepancies were mostly due to incomplete data annotation or specification of data processing and analysis.

- ★ Data made available through investigator's URL
  - ★ URL does not resolve (i.e., "rotten")

PLOS ONE | DOI:10.1371/journal.pone.0115253 December 26, 2014

RESEARCH ARTICLE

## Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot

Martin Klein<sup>1\*</sup>, Herbert Van de Sompel<sup>1</sup>, Robert Sanderson<sup>1</sup>, Harihar Shankar<sup>1</sup>, Lyudmila Balakireva<sup>1</sup>, Ke Zhou<sup>2</sup>, Richard Tobin<sup>2</sup>

We analyze a vast collection of articles from three corpora that span publication years 1997 to 2012. For over one million references to web resources extracted from over 3.5 million articles, we observe that the fraction of articles containing references to web resources is growing steadily over time. We find one out of five STM articles suffering from reference rot, meaning it is impossible to revisit the web context that surrounds them some time after their publication. When only considering STM articles that contain references to web resources, this fraction increases to seven out of ten.

# Better Approaches

## ★ Data paper

Ecological Research  
July 2013, Volume 28, Issue 4, p 541

Date: 10 May 2013

### Monitoring records of plant species in the Hakone region of Fuji-Hakone-Izu National Park, Japan, 2001–2010

Takeshi Osawa



### Abstract

The monitoring of species occurrences is a crucial aspect of biodiversity conservation, and regional volunteerism can serve as a powerful tool in such endeavors. The Fuji-Hakone-Izu National Park in the Hakone region of Kanagawa Prefecture, Japan, boasts a volunteer association of approximately 100 members. These volunteers have monitored plant species occurrences from 2001 to the present along several hiking trails in the region. In this paper, I present the annual observation records of plant occurrences in Hakone from 2001 to 2010. This data set includes 1,071 species of plants from 151 families. Scientific names follow the Y List, and this data set includes several threatened plant species. Data files are formatted based on the Darwin Core and Darwin Core Archives, which are defined by the Biodiversity Information Standards (BIS) or Biodiversity Information Standards Taxonomic Databases Working Group (TDWG). Data files filled on required and some additional item on Darwin Core. The data set can download from the author's personal Web site as of July 2012. These data will soon be published for the Global Biodiversity Information Facility (GBIF) through GBIF Japan. All users can then access the data from the GBIF portal site.

- The complete data set for this abstract published in the Data Paper section of the journal is available in electronic format in Ecological Research Data Paper Archives at [http://db.cger.nies.go.jp/JaLTER/ER\\_DataPapers/archives/2013/ERDP-2013-01](http://db.cger.nies.go.jp/JaLTER/ER_DataPapers/archives/2013/ERDP-2013-01).

## ★ Data published in a repository



The screenshot shows the LTER network interface. At the top, there's a logo for LTER and another for "The US Long Term Ecological Research Network". Below that, a search bar contains the text "WDNR Yahara Lakes Fisheries: Fish Lengths and Weights 1987-1998 - Lathrop". To the left of the search bar are buttons for "+/-" and "NTL LTER". To the right, there's a small orange dot icon. The main area displays the dataset details: "LTER Identifier: knb-lter-ntl.279.1" and "Abstract: These data were collected by the Wisconsin Department of Natural Resources (WDNR) from 1987-1998. Most of these data (1987-1993) precede 1995, the year that the University of Wisconsin took over sampling of the Yahara Lakes. However, WDNR data collected from 1997-1998 (unrelated to LTER sampling) is also included. In 1987 a joint project by the WDNR and the University of Wisconsin-Madison, Center for Limnology (CFL) was initiated on Lake Mendota. The project involved biomanipulation o...".

#### LTER Identifier:

knb-lter-ntl.279.1

#### Abstract:

These data were collected by the Wisconsin Department of Natural Resources (WDNR) from 1987-1998. Most of these data (1987-1993) precede 1995, the year that the University of Wisconsin took over sampling of the Yahara Lakes. However, WDNR data collected from 1997-1998 (unrelated to LTER sampling) is also included. In 1987 a joint project by the WDNR and the University of Wisconsin-Madison, Center for Limnology (CFL) was initiated on Lake Mendota. The project involved biomanipulation o...

#### Owners/Creators:

Lathrop

#### Metadata:

Select [here](#) for full metadata

#### Data File(s):

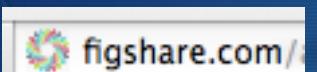
- [wdnr\\_fyke\\_minifyke\\_seine\\_lengths\\_weights.csv](#)
- [wdnr\\_boomshock\\_lengths\\_weights.csv](#)
- [wdnr\\_gillnet\\_lengths\\_weights\\_93.csv](#)
- [wdnr\\_walleye\\_age\\_lengths\\_weights\\_87.csv](#)
- [wdnr\\_creek\\_survey\\_lengths\\_weights.csv](#)
- [wdnr\\_creek\\_survey\\_angler\\_counts.csv](#)



# Goals of this Section

1. Understand best practices
2. Understand how to implement those best practices

# Making Data Accessible: Overview of Best Practices



## Highly connected drug file

Tretinoïn	257	46	Rv1155, aroG, Rv1264, mscL, thyX, gmk, glnA1, Rvi icl, Rv1264, thyX, glnA1, trpD, leuA, blaI, ethR
Levothyroxine	173	36	Rv0223c, lipJ, Rv1264, ephG, blaI, ethR, sigC, bj
Methotrexate	156	32	25 cyp130, Rv1264, lppX, gpm1, ligA, nirA
4-Hydroxytamoxifen	115	20	TB31.7, Rv1264, mscL, lppX, coaA, pcaA, Rv3676, fabG1,
Estradiol	98	1	mmaA4, bphD, Rv1264, mscL, thyX, lppX, mmaA2, pts TB31.7, cyp130, aroG, Rv1264, secA1, trpD, nirA
Amantadine	79		pth, ethR, clpP, glbN, inhA, pknD, lipJ, fabH, Rv1941, Rv3361c, Rv1264, lppX
Rifampin	78	13	mmaA4, Rv1264, groEL, lppX, secA1, glmU, nusA, R mvaA4, Rv1264, thyX, lppX, secA1, serA1, Rv3529, pepD, Rv1264, thyX, ethR, trxB2,
Raloxifene	75	18	pknD, pepD, fabH, Rv1941, devB, ppp, ftsZ, cyp12:
Propofol	54	5	
Indinavir	51	14	
Penicillamine	44	10	
Daunorubicin	44	12	
Triclosan	42	5	
Darunavir	40	15	

Enlarge to see the rest of the document

Enlarge Download

Cite this: Garijo, Daniel; Xie, Lei; Zhang, Yinliang; Gil, Yolanda; Xie, Li; Kinnings, Sarah;  
Bourne, Phil (2013): Highly connected drug file. figshare.  
<http://dx.doi.org/10.6084/m9.figshare.776887>  
Retrieved 08:56, Feb 20, 2015 (GMT)

## Description

Highly connected drug file obtained as a result of the TB-Drugome Workflow.

## Links

- <http://purl.org/net/tb-drugome-run>

1 Publication in a shared repository

2 General & domain metadata

3 Accessibility of Data (manual & machine)

4 Unique persistent identifier (PID)

5 Citation preference



# Best Practices (1 of 5)



## Highly connected drug file

Tretinoin	257	46	Rv1155, aroG, Rv1264, mscL, thyX, gmk, glnA1, Rvi icl, Rv1264, thyX, glnA1, trpD, leuA, blaI, ethR
Levothyroxine	173	36	Rv0223c, lipJ, Rv1264, ephG, blaI, ethR, sigC, bj
Methotrexate	156	32	25 cyp130, Rv1264, lppX, gpml, ligA, nirA
4-Hydroxytamoxifen	115	11	TB31.7, Rv1264, mscL, lppX, coaA, pcaA, Rv3676, e
Estradiol	98	20	fabG1,
Amantadine	79	1	mmaA4, bphD, Rv1264, mscL, thyX, lppX, mmaA2, pth
Rifampin	78	13	TB31.7, cyp130, aroG, Rv1264, secA1, trpD, nirA
Raloxifene	75	18	propofol, clpP, glbN, inha,
Propofol	54	5	pknD, lipJ, fabH, Rv1941, Rv3361c, Rv1264, lppX
Indinavir	51	14	mmaA4, Rv1264, groEL, lppX, secA1, glmU, nusA, Rv
Penicillamine	44	10	mmaA4, Rv1264, thyX, lppX, secA1, serA1, Rv3529c
Daunorubicin	44	12	pepD, Rv1264, thyX, ethR, trxB2,
Triclosan	42	5	pknD, pepD, fabH, Rv1941, devB, ppp, ftsZ, cyp12:
Darunavir	40	15	

Enlarge to see the rest of the document

Enlarge

Download

Cite this: Gario, Daniel; Xie, Lei; Zhang, Yinliang; Gil, Yolanda; Xie, Li; Kinnings, Sarah; Bourne, Phil (2013): Highly connected drug file. figshare.  
<http://dx.doi.org/10.6084/m9.figshare.776887>  
Retrieved 08:56, Feb 20, 2015 (GMT)

Published on 20 Aug 2013 - 12:44 (GMT)

Filesize is 4.96 KB

### Categories

- Computational Biology

### Authors

Daniel Gario  
Lei Xie  
Yinliang Zhang  
Yolanda Gil  
Li Xie  
Sarah Kinnings  
Phil Bourne

### Tags

- results
- tb-drugome

### LICENSE (what's this?)

CC-BY



1 Publication in a shared repository

2 General & domain metadata

3 Accessibility of Data (manual & machine)

4 Unique persistent identifier (PID)

5 Citation preference

## Description

Highly connected drug file obtained as a result of the TB-Drugome Workflow.

## Links

- <http://purl.org/net/tb-drugome-run>

# Popular Data Repositories

Not Curated

zenodo



zenodo  
figshare

The  
**Dataverse**  
Project



Curated

DRYAD



"Pangaea logo hg" by Hannes Grobe/AWI - Own work. Licensed under CC BY 3.0 via Wikimedia Commons - [http://commons.wikimedia.org/wiki/File:Pangaea\\_logo\\_hg.png#mediaviewer/File:Pangaea\\_logo\\_hg.png](http://commons.wikimedia.org/wiki/File:Pangaea_logo_hg.png#mediaviewer/File:Pangaea_logo_hg.png)

# Research Data Repositories

- <http://www.re3data.org>
- [http://databib.org/index\\_subjects.php](http://databib.org/index_subjects.php)
- [http://oad.simmons.edu/oadwiki/Data\\_repositories](http://oad.simmons.edu/oadwiki/Data_repositories)
- <http://www.force11.org>
- <http://www.nature.com/sdata/data-policies/repositories>



# International Geo Sample Number: IGSN

- ★ Globally unique and persistent identifier for physical samples in the Earth Sciences
- ★ Obtain IGSNs for your samples
  - ★ Best upon collection or as soon as you are back online!
- ★ Go to <http://www.geosamples.org/> or contact info@geosamples.org
- ★ Record and register quality metadata for your samples
  - ★ At a minimum: Location, Lithology, Contact, access restrictions
- ★ Use IGSNs in your publications: text, data tables,...

IGSN: GMY00007W



IGSN: GMY00007W  
Sample Name: TN182\_47\_002  
Other Name(s):  
Sample Type: Individual Sample  
Parent IGSN: GMY00001B

## Description

Material:	Rock
Classification:	Igneous>Plutonic>Mafic
Field Name:	gabbro, hornblende gabbro
Description:	mafic plutonic rock

# Best Practices (2 of 5)



## Highly connected drug file

Tretinoin	257	46	Rv1155, aroG, Rv1264, mscL, thyX, gmk, glnA1, Rvi icl, Rv1264, thyX, glnA1, trpD, leuA, blaI, ethR
Levothyroxine	173	36	Rv0223c, lipJ, Rv1264, ephG, blaI, ethR, sigC, bj
Methotrexate	156	32	25 cyp130, Rv1264, lppX, gpml, ligA, nirA
4-Hydroxytamoxifen	115	11	TB31.7, Rv1264, mscL, lppX, coaA, pcaA, Rv3676, e
Estradiol	98	20	fabG1,
Amantadine	79	1	mmaA4, bphD, Rv1264, mscL, thyX, lppX, mmaA2, pti
Rifampin	78	13	TB31.7, cyp130, aroG, Rv1264, secA1, trpD, nirA
Raloxifene	75	18	pth, ethR, clpP, glbN, inha,
Propofol	54	5	pknD, lipJ, fabH, Rv1941, Rv3361c, Rv1264, lppX
Indinavir	51	14	mmaA4, Rv1264, groEL, lppX, secA1, glmU, nusA, Rv
Penicillamine	44	10	mmaA4, Rv1264, thyX, lppX, secA1, serA1, Rv3529c
Daunorubicin	44	12	pepD, Rv1264, thyX, ethR, trxB2,
Triclosan	42	5	pknD, pepD, fabH, Rv1941, devB, ppp, ftsZ, cyp12:
Darunavir	40	15	

Enlarge to see the rest of the document

Enlarge

Download

Cite this: Gario, Daniel; Xie, Lei; Zhang, Yinliang; Gil, Yolanda; Xie, Li; Kinnings, Sarah; Bourne, Phil (2013): Highly connected drug file. figshare.  
<http://dx.doi.org/10.6084/m9.figshare.776887>  
Retrieved 08:56, Feb 20, 2015 (GMT)

## Description

Highly connected drug file obtained as a result of the TB-Drugome Workflow.

## Links

- <http://purl.org/net/tb-drugome-run>

1

Publication in a shared repository

2

General & domain metadata

3

Accessibility of data (manual & machine)

4

Unique persistent identifier (PID)



5

Citation preference

# Minimal Metadata

## General

- ★ Dataset name/title
- ★ Description
- ★ Creator(s)
- ★ Publication date
- ★ License
- ★ Publisher/contact
- ★ Version
- ★ Resource type
- ★ Location of the data

Typical of digital libraries,  
eg the Dublin Core  
standard  
(<http://dublincore.org/documents/dcmi-terms/>)

# Choose a License

Screenshot of the Creative Commons License Chooser interface:

**License Features**  
Your choices on this panel will update the other panels on this page.

Allow adaptations of your work to be shared?  
 Yes  No  
 Yes, as long as others share alike

Allow commercial uses of your work?  
 Yes  No

**Selected License**  
**Attribution 4.0 International**

This is a Free Culture License!



**Help others attribute you!**  
This part is optional, but filling it out will add machine-readable metadata to the suggested HTML!

Title of work   
Attribute work to name   
Attribute work to URL   
Source work URL   
More permissions URL   
Format of work   
License mark

**Have a web page?**

This work is licensed under a Creative Commons Attribution 4.0 International License.

Copy this code to let your visitors know!

```
<a rel="license" href="http://creativecommons.org/licenses/by/4.0/"></a><br />This work is licensed under a <a rel="license" href="http://creativecommons.org/licenses/by/4.0/">Creative Commons Attribution 4.0 International License</a>
```

Normal Icon  Compact Icon

## Recommended: CC-BY and CC0



### Attribution CC BY

This license lets others distribute, remix, tweak, and build upon your work, even commercially, as long as they credit you for the original creation. This is the most accommodating of licenses offered. Recommended for maximum dissemination and use of licensed materials.

### CC0 (datasets) “No rights reserved”



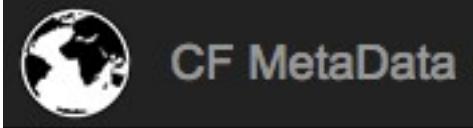
CC0 can be particularly important for the sharing of data and databases, since it otherwise may be unclear whether highly factual data and databases are restricted by copyright or other rights.

Databases may contain facts that, in and of themselves, are not protected by copyright law.

CC0 is recommended for data and databases and is used by hundreds of organizations. It is especially recommended for scientific data. Although CC0 doesn't legally require users of the data to cite the source, it does not take away the moral responsibility to give attribution, as is common in scientific research.

<http://creativecommons.org/licenses/>

# Domain-Specific Metadata Standards



ISO 19115

Marine Metadata Interoperability

WaterML2.0

- ★ A data repository in a given discipline may request metadata using accepted standards

# Best Practices (3 of 5)



## Highly connected drug file

Tretinoin	257	46	Rv1155, aroG, Rv1264, mscL, thyX, gmk, glnA1, Rvi icl, Rv1264, thyX, glnA1, trpD, leuA, blaI, ethR
Levothyroxine	173	36	Rv0223c, lipJ, Rv1264, ephG, blaI, ethR, sigC, bj
Methotrexate	156	32	25 cyp130, Rv1264, lppX, gpml, ligA, nirA
4-Hydroxytamoxifen	115	11	TB31.7, Rv1264, mscL, lppX, coaA, pcaA, Rv3676, e
Estradiol	98	20	fabG1,
Amantadine	79	1	mmaA4, bphD, Rv1264, mscL, thyX, lppX, mmaA2, pti
Rifampin	78	13	TB31.7, cyp130, aroG, Rv1264, secA1, trpD, nirA
Raloxifene	75	18	pth, ethR, clpP, glbN, inha,
Propofol	54	5	pknD, lipJ, fabH, Rv1941, Rv3361c, Rv1264, lppX
Indinavir	51	14	mmaA4, Rv1264, groEL, lppX, secA1, glmU, nusA, Rv
Penicillamine	44	10	mmaA4, Rv1264, thyX, lppX, secA1, serA1, Rv3529c
Daunorubicin	44	12	pepD, Rv1264, thyX, ethR, trxB2,
Triclosan	42	5	pknD, pepD, fabH, Rv1941, devB, ppp, ftsZ, cyp12:
Darunavir	40	15	

Enlarge to see the rest of the document

Enlarge

Download

Cite this:

Garijo, Daniel; Xie, Lei; Zhang, Yinliang; Gil, Yolanda; Xie, Li; Kinnings, Sarah;

Bourne, Phil (2013): Highly connected drug file. figshare.

<http://dx.doi.org/10.6084/m9.figshare.776887>

Retrieved 08:56, Feb 20, 2015 (GMT)

## Description

Highly connected drug file obtained as a result of the TB-Drugome Workflow.

## Links

- <http://purl.org/net/tb-drugome-run>

1

Publication in a shared repository

Published on 20 Aug 2013 - 12:44 (GMT)

Filesize is 4.96 KB

### Categories

- Computational Biology

### Authors

Daniel Garijo

Lei Xie

Yinliang Zhang

Yolanda Gil

Li Xie

Sarah Kinnings

Phil Bourne

### Tags

- results
- tb-drugome

### LICENSE (what's this?)

CC-BY



2

General & domain metadata

3

Accessibility of data (manual & machine)

4

Unique persistent identifier (PID)

5

Citation preference

# Manual Accessibility

## UNIQUE ID & METADATA

- ★ [http://figshare.com/articles/Highly\\_connected\\_drug\\_file/776887](http://figshare.com/articles/Highly_connected_drug_file/776887)

A screenshot of a figshare article page. The title is "Highly connected drug file". The page shows a table of highly connected drugs with their counts and IDs. Below the table, there is a "Download" button. A large green arrow points from the top of the slide towards this "Download" button.

Drug	Count	ID
Tretinoin	257	Rv1155, aroG, Rv1264, mscL, thyX, glnA1, Rv1264, thyX, glnA1, trpD, leuA, blaI, ethR, Rv0223c, lipJ, Rv1264, cyp130, Rv1264
Estradiol	98	TB31.7, Rv1264, mscL, lipJ, fabG1, Rv1264
Amantadine	79	cyp130, Rv1264, mscL, lipJ, fabG1, Rv1264
Rifampin	78	mmaA4, bphD, Rv1264, mscL, lipJ, fabG1, Rv1264
Raloxifene	75	cyp130, Rv1264, mscL, lipJ, fabG1, Rv1264
Propofol	54	pth, ethR, clpP, glbN, inhA, pknD, lipJ, fabH, Rv1941, Rv3361c, Rv1264
Indinavir	51	mmaA4, Rv1264, groEL, lppX, secA1, glmU, nusA, Rv1264
Penicillamine	44	10
Daunorubicin	44	12
Triclosan	42	5
Darunavir	40	15

Below the table, there is a "Download" button. The page also includes sections for "Share this", "Cite this", "Description", "Links", and "Comments".

## DATA

- ★ <http://files.figshare.com/1175525/highConnectedDrugs.txt>

A screenshot of a file listing page. It shows a table of highly connected drugs with their counts and IDs. A large green arrow points from the top of the slide towards this table.

Tretinoin	257	46	Rv1155, aroG, Rv1264, mscL, thyX, glnA1, Rv1264, thyX, glnA1, trpD, leuA, blaI, ethR, Rv0223c, lipJ, Rv1264, cyp130, Rv1264
Levothyroxine	173	36	TB31.7, Rv1264, mscL, lipJ, fabG1, Rv1264
Methotrexate	156	32	cyp130, Rv1264, lppX, gpml, ligA, nirA, TB31.7, Rv1264, mscL, lppX, coaA, pcaA, Rv3676, fabG1, mmaA4, bphD, Rv1264, mscL, thyX, lppX, mmaA2, pt, TB31.7, cyp130, aroG, Rv1264, secA1, trpD, nirA, pth, ethR, clpP, glbN, inhA, pknD, lipJ, fabH, Rv1941, Rv3361c, Rv1264, lppX, mmaA4, Rv1264, groEL, lppX, secA1, glmU, nusA, Rv1264, mmaA4, Rv1264, thyX, lppX, secA1, serA1, Rv3529, pepD, Rv1264, thyX, ethR, trx2B, pknD, pepD, fabH, Rv1941, devB, ppp, ftsZ, cyp12
4-Hydroxytamoxifen	115	25	cyp130, Rv1264, lppX, gpml, ligA, nirA, TB31.7, Rv1264, mscL, lppX, coaA, pcaA, Rv3676, fabG1, mmaA4, bphD, Rv1264, mscL, thyX, lppX, mmaA2, pt, TB31.7, cyp130, aroG, Rv1264, secA1, trpD, nirA, pth, ethR, clpP, glbN, inhA, pknD, lipJ, fabH, Rv1941, Rv3361c, Rv1264, lppX, mmaA4, Rv1264, groEL, lppX, secA1, glmU, nusA, Rv1264, mmaA4, Rv1264, thyX, lppX, secA1, serA1, Rv3529, pepD, Rv1264, thyX, ethR, trx2B, pknD, pepD, fabH, Rv1941, devB, ppp, ftsZ, cyp12
Estradiol	98	20	TB31.7, Rv1264, mscL, lipJ, fabG1, Rv1264
Amantadine	79	1	cyp130, Rv1264, mscL, lipJ, fabG1, Rv1264
Rifampin	78	13	mmaA4, bphD, Rv1264, mscL, thyX, lppX, mmaA2, pt, TB31.7, cyp130, aroG, Rv1264, secA1, trpD, nirA, pth, ethR, clpP, glbN, inhA, pknD, lipJ, fabH, Rv1941, Rv3361c, Rv1264, lppX, mmaA4, Rv1264, groEL, lppX, secA1, glmU, nusA, Rv1264, mmaA4, Rv1264, thyX, lppX, secA1, serA1, Rv3529, pepD, Rv1264, thyX, ethR, trx2B, pknD, pepD, fabH, Rv1941, devB, ppp, ftsZ, cyp12
Raloxifene	75	18	cyp130, Rv1264, mscL, lipJ, fabG1, Rv1264
Propofol	54	5	pth, ethR, clpP, glbN, inhA, pknD, lipJ, fabH, Rv1941, Rv3361c, Rv1264, lppX, mmaA4, Rv1264, groEL, lppX, secA1, glmU, nusA, Rv1264, mmaA4, Rv1264, thyX, lppX, secA1, serA1, Rv3529, pepD, Rv1264, thyX, ethR, trx2B, pknD, pepD, fabH, Rv1941, devB, ppp, ftsZ, cyp12
Indinavir	51	14	pknD, lipJ, fabH, Rv1941, Rv3361c, Rv1264, lppX, mmaA4, Rv1264, groEL, lppX, secA1, glmU, nusA, Rv1264, mmaA4, Rv1264, thyX, lppX, secA1, serA1, Rv3529, pepD, Rv1264, thyX, ethR, trx2B, pknD, pepD, fabH, Rv1941, devB, ppp, ftsZ, cyp12
Penicillamine	44	10	cyp130, Rv1264, lppX, gpml, ligA, nirA, TB31.7, Rv1264, mscL, lppX, coaA, pcaA, Rv3676, fabG1, mmaA4, Rv1264, thyX, lppX, secA1, serA1, Rv3529, pepD, Rv1264, thyX, ethR, trx2B, pknD, pepD, fabH, Rv1941, devB, ppp, ftsZ, cyp12
Daunorubicin	44	12	cyp130, Rv1264, lppX, gpml, ligA, nirA, TB31.7, Rv1264, mscL, lppX, coaA, pcaA, Rv3676, fabG1, mmaA4, Rv1264, thyX, lppX, secA1, serA1, Rv3529, pepD, Rv1264, thyX, ethR, trx2B, pknD, pepD, fabH, Rv1941, devB, ppp, ftsZ, cyp12
Triclosan	42	5	cyp130, Rv1264, lppX, gpml, ligA, nirA, TB31.7, Rv1264, mscL, lppX, coaA, pcaA, Rv3676, fabG1, mmaA4, Rv1264, thyX, lppX, secA1, serA1, Rv3529, pepD, Rv1264, thyX, ethR, trx2B, pknD, pepD, fabH, Rv1941, devB, ppp, ftsZ, cyp12
Darunavir	40	15	cyp130, Rv1264, lppX, gpml, ligA, nirA, TB31.7, Rv1264, mscL, lppX, coaA, pcaA, Rv3676, fabG1, mmaA4, Rv1264, thyX, lppX, secA1, serA1, Rv3529, pepD, Rv1264, thyX, ethR, trx2B, pknD, pepD, fabH, Rv1941, devB, ppp, ftsZ, cyp12

# Machine Accessibility: Metadata is a Necessity!

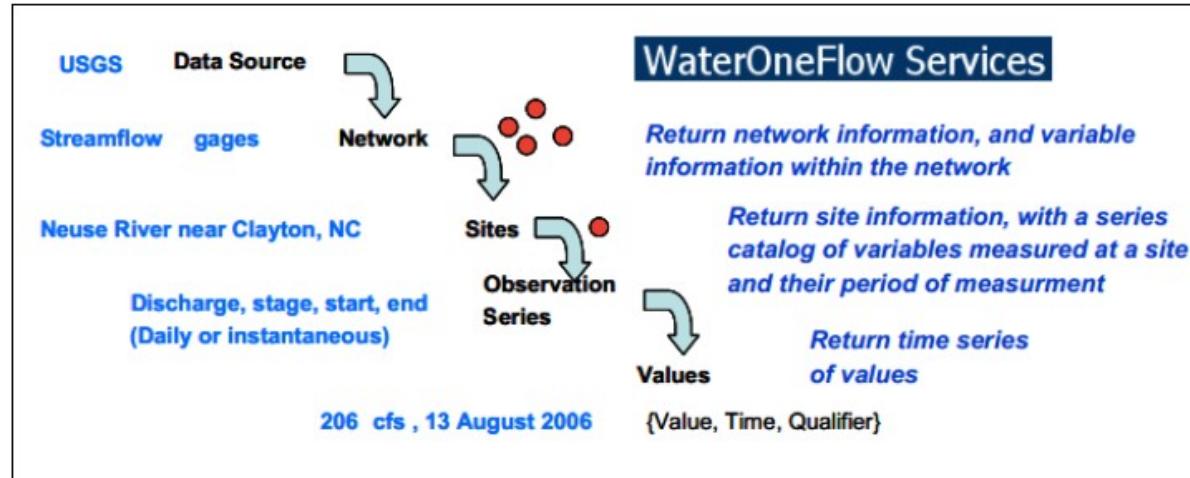


# CUAHSI

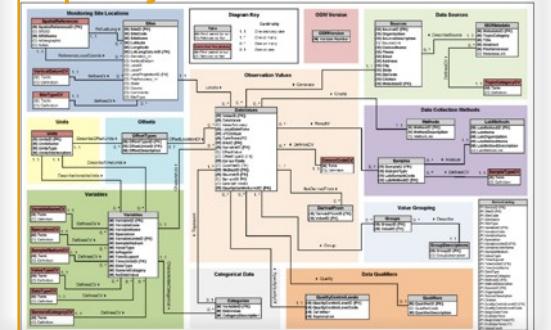
UNIVERSITIES ALLIED FOR WATER RESEARCH

## WaterOneFlow Web Services

Web services are computer applications that interact with and exchange information with other applications over the internet. The CUAHSI HIS uses a family of web services, called WaterOneFlow (WOF), that have been developed as a standard mechanism for the transfer of hydrologic data between hydrologic data servers (databases) and users' computers. Web services streamline the often time consuming tasks of extracting data from a data source, transforming it into a usable format, and loading it in to an analysis environment. The WaterOneFlow Web Services format the data as the type of XML described above, WaterML 1.1.



Data model specifies how to query the data available



# Best Practices (4 of 5)



## Highly connected drug file

Tretinoin	257	46	Rv1155, aroG, Rv1264, mscL, thyX, gmk, glnA1, Rvi icl, Rv1264, thyX, glnA1, trpD, leuA, blaI, ethR
Levothyroxine	173	36	Rv0223c, lipJ, Rv1264, ephG, blaI, ethR, sigC, bj
Methotrexate	156	32	25 cyp130, Rv1264, lppX, gpml, ligA, nirA
4-Hydroxytamoxifen	115	11	TB31.7, Rv1264, mscL, lppX, coaA, pcaA, Rv3676, e
Estradiol	98	20	fabG1,
Amantadine	79	1	mmaA4, bphD, Rv1264, mscL, thyX, lppX, mmaA2, pti
Rifampin	78	13	TB31.7, cyp130, aroG, Rv1264, secA1, trpD, nirA
Raloxifene	75	18	pth, ethR, clpP, glbN, inha,
Propofol	54	5	pknD, lipJ, fabH, Rv1941, Rv3361c, Rv1264, lppX
Indinavir	51	14	mmaA4, Rv1264, groEL, lppX, secA1, glmU, nusA, Rv
Penicillamine	44	10	mmaA4, Rv1264, thyX, lppX, secA1, serA1, Rv3529c
Daunorubicin	44	12	pepD, Rv1264, thyX, ethR, trxB2,
Triclosan	42	5	pknD, pepD, fabH, Rv1941, devB, ppp, ftsZ, cyp12:
Darunavir	40	15	

Enlarge to see the rest of the document

Enlarge

Download

Cite this: Gario, Daniel; Xie, Lei; Zhang, Yinliang; Gil, Yolanda; Xie, Li; Kinnings, Sarah; Bourne, Phil (2013): Highly connected drug file. figshare.  
<http://dx.doi.org/10.6084/m9.figshare.776887>  
Retrieved 08:56, Feb 20, 2015 (GMT)

## Description

Highly connected drug file obtained as a result of the TB-Drugome Workflow.

## Links

- <http://purl.org/net/tb-drugome-run>

1

Publication in a shared repository

Published on 20 Aug 2013 - 12:44 (GMT)

Filesize is 4.96 KB

### Categories

- Computational Biology

### Authors

Daniel Gario

Lei Xie

Yinliang Zhang

Yolanda Gil

Li Xie

Sarah Kinnings

Phil Bourne

### Tags

- results
- tb-drugome

### LICENSE (what's this?)

CC-BY



2

General & domain metadata

3

Accessibility of data (manual & machine)

4

Unique persistent identifier (PID)

5

Citation preference

# Main Types of Unique Identifiers

1. Uniform Resource Locator (URL)
2. Persistent URL (PURL)
3. Digital Object Identifier



# URL/URI

- Minimal effort to create
- No guarantee of persistence
  - i.e., almost guaranteed it will not have persistence
  - e.g.,  
`http://www.greatuniversity.edu/gradstudents/joesmith/awesomedata/`

**Do not use in papers!!**

# Persistent URL (PURL)



- The same PURL can be resolved to different Web address over time
  - You always refer to your data with the same PURL:  
<http://purl.org/mydataandme/awesomedata.html>
    - Today you are in grad school and tell purl.org to resolve it to:  
[http://www.wisc.edu/myadvisorsgroup/a\\_wesomedata.html](http://www.wisc.edu/myadvisorsgroup/a_wesomedata.html)
    - Tomorrow you have graduated and tell purl.org to resolve it to:  
[http://www.stanford.edu/myowngroup/a\\_wesomedata.html](http://www.stanford.edu/myowngroup/a_wesomedata.html)
  - It is easy to create your own PURLs, just remember to update whenever you move the data
    - Go to <https://w3id.org> (run by W3C), <http://www.purl.org> (run by OCLC), or other PURL services

# Digital Object Identifier (DOI)

PLoS Biol. 2003 Nov; 1(2): e57.

Published online 2003 Nov 17 doi: [10.1371/journal.pbio.0000057](https://doi.org/10.1371/journal.pbio.0000057)

## The What and Whys of DOIs

[Susanne DeRisi](#), [Rebecca Kennison](#), and [Nick Twyman](#)

[Copyright and License information ►](#)

This article has been [cited by other articles in PMC](#).

DOIs can only be issued by a DOI authority (eg a journal publisher) that guarantees to always resolve to a permanent location.

Data repositories can issue DOIs for data

DOIs are free

As you may have noticed in the first issue of *PLoS Biology* and again in this issue, there are many places where an alphanumeric string appears after the letters “DOI,” such as [10.1371/journal.pbio.000005](https://doi.org/10.1371/journal.pbio.000005) or [10.1371/journal.pbio.000005.g005](https://doi.org/10.1371/journal.pbio.000005.g005). Although some of you may already be acquainted with DOIs, others of you may wonder what they are, how they are used, and why we are using them.

## What Are DOIs?

Go to:

A Digital Object Identifier (DOI) is an URN (Uniform Resource Name), a compact string that provides a unique, persistent, and actionable identifier for the digital object with which it is associated. DOIs are commonly assigned to scientific articles in their electronic form, but DOIs may also be used as identifiers for any object in any location, although this usage is not yet common outside the online world. The International DOI Foundation (IDF), which governs the DOI system, has several hundred registrant organizations and in August 2003 reported that over 10 million DOIs have been issued since the foundation was created in 1998 (<http://www.doi.org/news/03augnews.html>).

# Best Practices (5 of 5)



Highly connected drug file

Tretinoin	257	46	Rv1155, aroG, Rv1264, mscL, thyX, gmk, glnA1, Rvi icl, Rv1264, thyX, glnA1, trpD, leuA, blaI, ethR
Levothyroxine	173	36	Rv0223c, lipJ, Rv1264, ephG, blaI, ethR, sigC, bj
Methotrexate	156	32	25 cyp130, Rv1264, lppX, gpml, ligA, nirA
4-Hydroxytamoxifen	115	11	TB31.7, Rv1264, mscL, lppX, coaA, pcaA, Rv3676, e
Estradiol	98	20	fabG1,
Amantadine	79	1	mmaA4, bphD, Rv1264, mscL, thyX, lppX, mmaA2, pti
Rifampin	78	13	TB31.7, cyp130, aroG, Rv1264, secA1, trpD, nirA
Raloxifene	75	18	pth, ethR, clpP, glbN, inha,
Propofol	54	5	pknD, lipJ, fabH, Rv1941, Rv3361c, Rv1264, lppX
Indinavir	51	14	mmaA4, Rv1264, groEL, lppX, secA1, glmU, nusA, Rv
Penicillamine	44	10	mmaA4, Rv1264, thyX, lppX, secA1, serA1, Rv3529c
Daunorubicin	44	12	pepD, Rv1264, thyX, ethR, trxB2,
Triclosan	42	5	pknD, pepD, fabH, Rv1941, devB, ppp, ftsZ, cyp12:
Darunavir	40	15	

Enlarge to see the rest of the document

Enlarge

Download

Cite this: Gario, Daniel; Xie, Lei; Zhang, Yinliang; Gil, Yolanda; Xie, Li; Kinnings, Sarah; Bourne, Phil (2013): Highly connected drug file. figshare.  
<http://dx.doi.org/10.6084/m9.figshare.776887>  
Retrieved 08:56, Feb 20, 2015 (GMT)

Published on 20 Aug 2013 - 12:44 (GMT)

Filesize is 4.96 KB

## Categories

- Computational Biology

## Authors

Daniel Gario

Lei Xie

Yinliang Zhang

Yolanda Gil

Li Xie

Sarah Kinnings

Phil Bourne

## Tags

- results
- tb-drugome

## LICENSE (what's this?)

CC-BY



1

Publication in a shared repository

2

General & domain metadata

3

Accessibility of data (manual & machine)

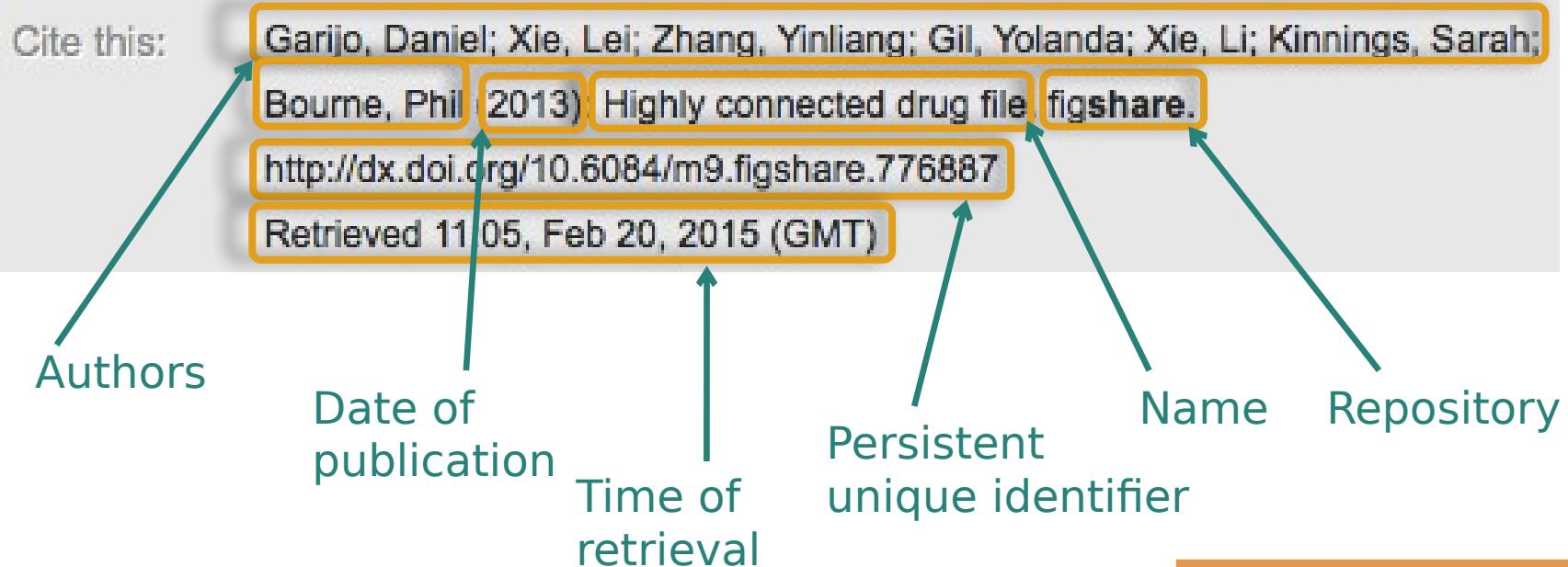
4

Unique persistent identifier (PID)

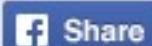
5

Citation preference

# Data Citation Format



Share this:



0



0



Embed\*

Data repositories and journals often specify how to cite data



# Goals of this Section

1. Understand what those best practices mean
2. Understand how to implement those best practices

# Accessible. Simplest Approach

1

zenodo

2



3

Rv1155, aroG, aroI, Rv1264, thyR, Rv0223c, lipJ, Rv125, cyp130, Rv1264, TB31.7, Rv1264, mscL, fabG1, mmaA4, bphD, Rv1264, mmaA4, TB31.7, cyp130, aroG, pth, ethR, clpP, glbN, pknD, lipJ, fabH, Rv1264, mmaA4, Rv1264, groEI, mmaA4, Rv1264, thyR, pepD, Rv1264, thyR, pknD, pepD, fabH

## 1. Create a public entry for your dataset with a persistent unique identifier

- Go to a domain repository (use a general repository, e.g., zenodo.org, if you cannot find one), create an account
- Create an entry for your dataset

## 2. Specify the metadata

- Including license -- choose from <http://www.creativecommons.org/licenses>

## 3. Upload/point to the data

**Voilà! The repository will give**

# Making Data Accessible: Ideal Approach



1. Find a repository that your community uses, if there is not one then organize one!
2. Create a public entry for your dataset with a persistent unique identifier
  - Create an entry for your dataset
3. Specify the metadata
  - Including license -- choose from <http://www.creativecommons.org/licenses>
4. Upload/point to the data
5. Get a data citation from the repository

## Making Data Accessible:

# Cite the data in your paper

Initial raw data

Intermediate data

Final data

- ★ **Citation goes in the References section**
- ★ **How to cite the data? You choose:**
  - ★ With an in-text pointer as you would cite any other paper (recommended)
  - ★ With an in-text pointer in a special “Data Resources” section
  - ★ With an in-text pointer in the “Acknowledgments” section

# Making Software Accessible

## OntoSoft Training

### Part 3

<http://dx.doi.org/10.5281/zenodo.15920>

<http://www.scientificpaperofthefuture.org>



CC-BY  
Attribution



# The Value of Software

## Availability of Software



PLOS supports the development of open source software and believes that, for submissions appropriate open source standards will ensure that the submission conforms to (1) our requirement that the software can be used by other researcher can reproduce the experiments described, (2) our aim to promote openness so that others can build upon the work, and (3) our desire that PLOS journals can be built upon by future researchers. Therefore, if new software or a new application is developed and deposited in a public archive, and it is determined that the software conforms to the [Open Source Definition](#), have deposited the following three items as Supporting Information:

- **The associated source code of the software described by the paper.** This should be licensed under a suitable license such as BSD, LGPL, or MIT (see <http://www.openaccessrepository.org>). The use of commercial software such as Mathematica and MATLAB does not preclude a paper from being open access, but it is preferred.
- **Documentation for running and installing the software.** For end-user applications, this may include instructions for installing prerequisites; for software libraries, instructions for using the application program interface.
- **A test dataset with associated control parameter settings.** Where feasible, results should be reproducible. Test data should not have any dependencies — for example, a database dump.

Acceptable archives should provide a public repository of the described software. The code can be deposited in a public archive such as SourceForge, GitHub, Bitbucket, or Google Code. Examples include the SourceForge project for creating user accounts, logging in, and publishing software, which contains more than 1,000 projects. Examples also include the SourceForge project for the Savannah, GitHub and the Codehaus.

**nature**

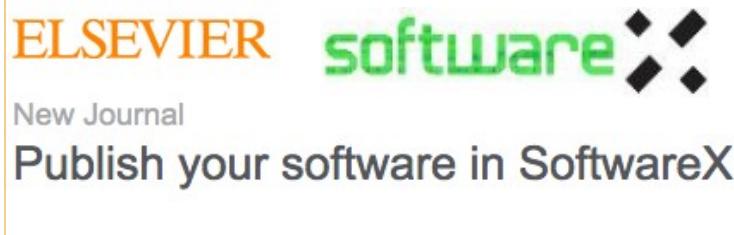
*Nature*, 467, pp 753, 2010.  
doi:10.1038/467753a

## Publish your computer code: it is good enough

*Freely provided working code — whatever its quality — improves programming and enables others to engage with your research, says Nick Barnes.*

# Software Papers and Software Repositories

- ★ Some journal articles describe a piece of software
- ★ Some publications have “software papers” or “software metapapers”



Apache Open Climate Workbench

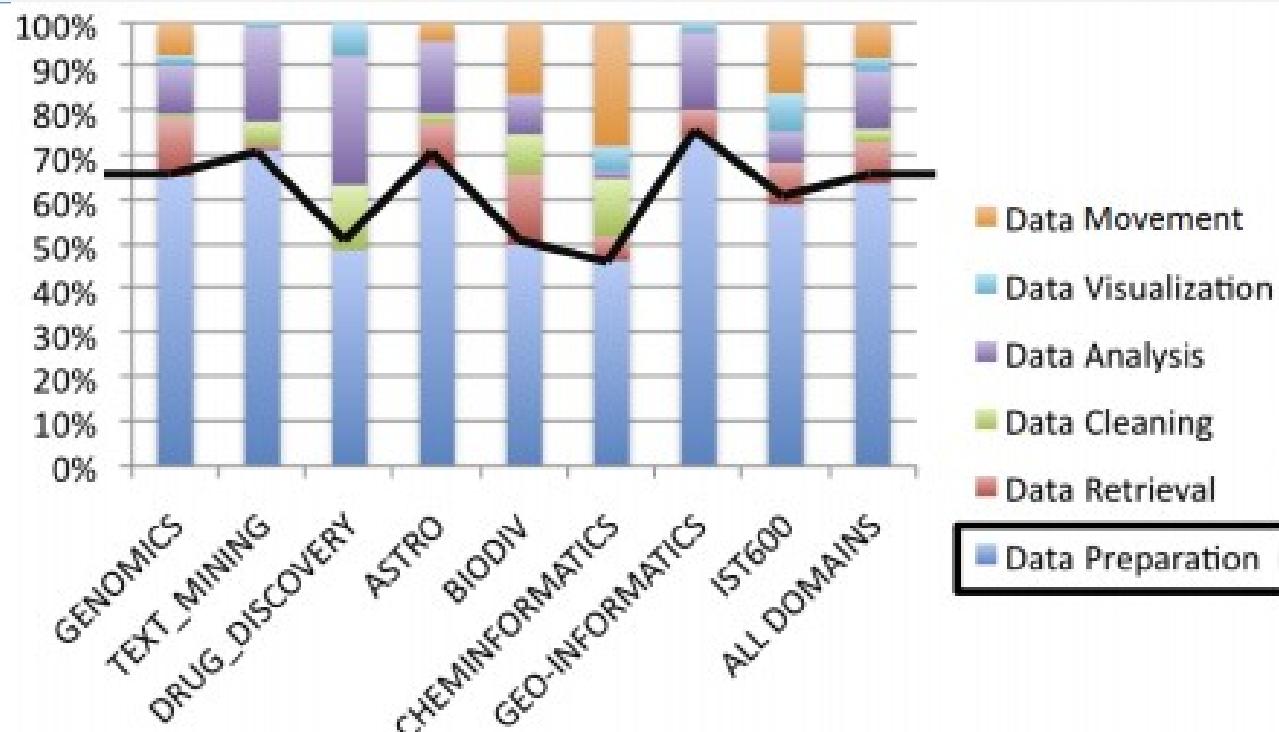


# Why Is Scientific Software Not Shared?

- ★ “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 commercialize my software”
- ★ “I don’t know where to start!”

# Data Preparation Software Dominates but is Least Shared

- ★ “Scientists and engineers spend more than 60% of their time just preparing the data for model input or data-model comparison” (NASA A40)



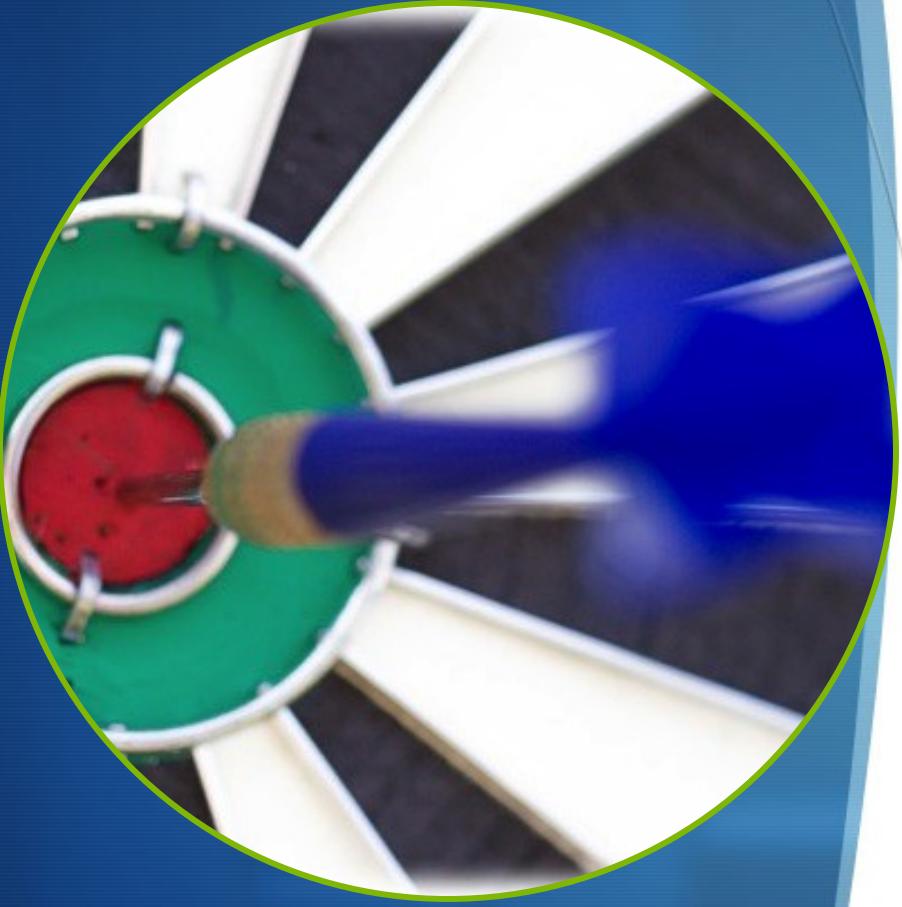
**“Common Motifs in Scientific Workflows: An Empirical Analysis.” Garijo, D.; Alper, P.; Belhajjame, K.; Corcho, O.; Gil, Y.; and Goble, C. Future Generation Computer Systems**

# “Dark Software”



- ★ Models that are not published
  - ★ Eg from a PhD thesis
- ★ Data preparation software
- ★ Visualization software

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



# Goals of this Section

1. Making software ready for publication
2. Understand best practices in software publication
3. Understand how to implement those best practices

# Best Practices



1. Accessible from a public location
2. License
3. Citation

# Making Software Accessible from a Public Location

**PURL**

zenodo

 GitHub

 The Apache Software Foundation  
Community-led development since 1999.

## Options:

- ★ **Publish in your web site**
  - ★ Very easy and simple
  - ★ Get a PURL for the version you use in the paper
- ★ **Use a data repository** (eg zenodo), treating code like data
  - ★ Very easy and simple
  - ★ It allows you to get a DOI
- ★ **Use a code repository** (eg GitHub, BitBucket)
  - ★ Beneficial if you have other users or want to track new versions
  - ★ Some will give you a DOI (eg GitHub)
- ★ **Create a formal community project** (eg in Apache)

# Choosing an Open Source License

- ★ Copyright: automatically applied to software when it is created to grant *the creator* exclusive rights as an intellectual property
- ★ **Open source license:** reduce constraints and enable software developers to make their source code available to public
  1. “Copyleft” license (ex: GNU General Public License (GPL))
  2. “Permissive” license (ex: Apache 2 or MIT licenses)
- ★ **Open Source Initiative**
  - ★ Choose a license from: <http://opensource.org/licenses>
  - ★ Recommend that you choose a permissive license
    - ★ Apache v2



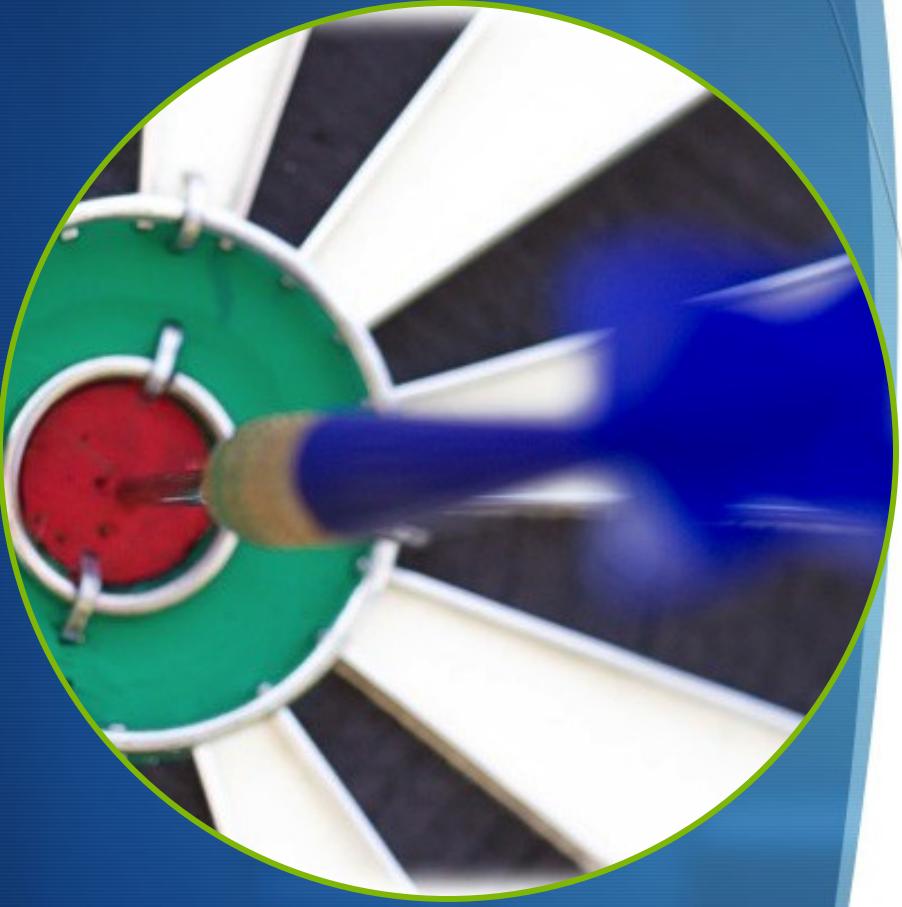
# Software Citation

- ★ Use a persistent unique identifier (PURL or DOI)
  - ★ Analogous to identifiers for data
- ★ Software sharing repositories are beginning to offer the ability to assign DOIs

# Software Citation Format

- ★ Similar to data citation format, but includes software version

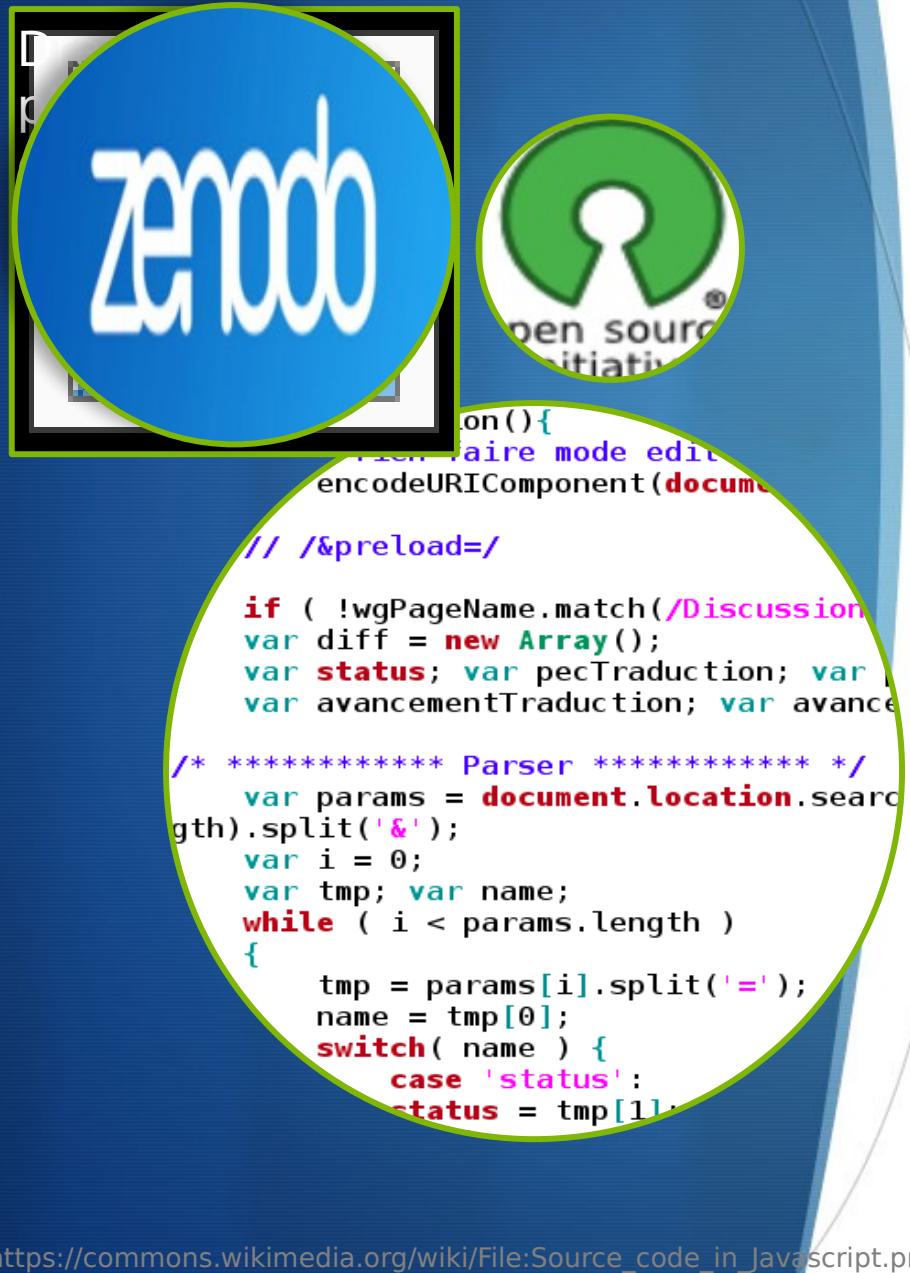




# Goals of this Section

1. Making software ready for publication
2. Understand best practices in software publication
3. Understand how to implement those best practices

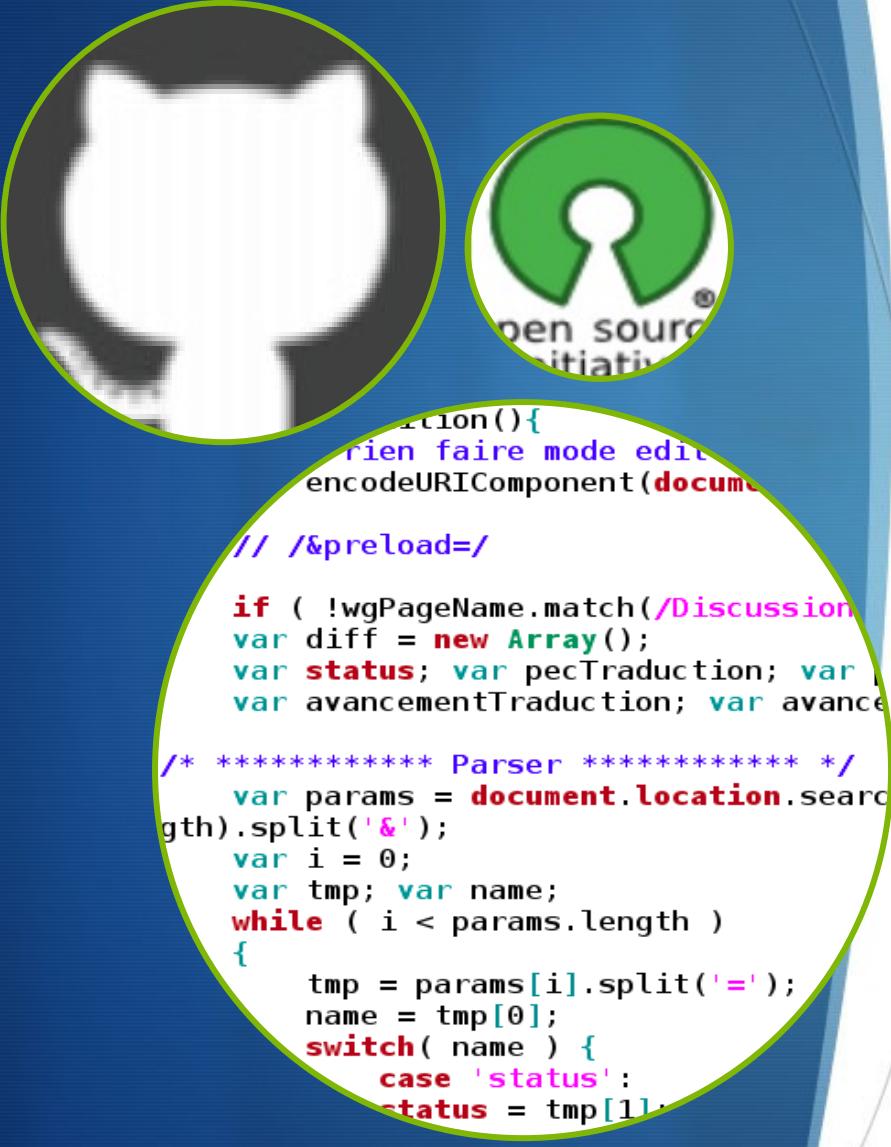
# Making Software Accessible: Simplest Approach



1. Create a public entry for your software with a persistent unique identifier
  - Upload to a data repository (e.g., Zenodo) as you would data, and get a DOI
    - Or post on your web site and use a PURL
2. Specify basic metadata
  - Including license -- choose from <http://opensource.org/licenses>, preferably Apache v2.0
3. Specify desired citation

Accessible.

# Ideal Approach



```
function(){  
    rien faire mode edit  
    encodeURIComponent(document.  
  
// /&preload=  
  
if ( !wgPageName.match(/Discussion/)  
var diff = new Array();  
var status; var pecTraduction; var p  
var avancementTraduction; var avance  
  
/* ***** Parser ***** */  
var params = document.location.search  
gth).split('&');  
var i = 0;  
var tmp; var name;  
while ( i < params.length )  
{  
    tmp = params[i].split('=');  
    name = tmp[0];  
    switch( name ) {  
        case 'status':  
            status = tmp[1];  
            break;  
    }  
}
```

1. Learn to use a code repository that allows version tracking and collaborative software development
  - GitHub, BitBucket, etc.
2. Create a public entry for your software with a persistent unique identifier
3. Specify the metadata
  - Including license -- choose from <http://opensource.org/licenses>, preferably Apache v2.0
4. Specify desired citation

## Making Software Accessible:

# Cite the software in your paper

Analogous to citing data:

- ★ Citation goes in the References section
- ★ How to cite the software?  
You choose:
  - ★ With an in-text pointer as you would cite any other paper (recommended)
  - ★ With an in-text pointer in a special “Data Resources” (or “Software Resources”) section
  - ★ With an in-text pointer in the “Acknowledgments” section

# Documenting Software through Metadata

## OntoSoft Training

### Part 4

<http://dx.doi.org/10.5281/zenodo.15920>



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

CC-BY  
Attribution



EarthCube



# Goals of this Section

1. Understand what metadata needs to be documented about software to promote reuse
2. Understand how to use a software registry to specify that metadata

# Software Repository vs Software Registry

## ★ **Software repository**

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

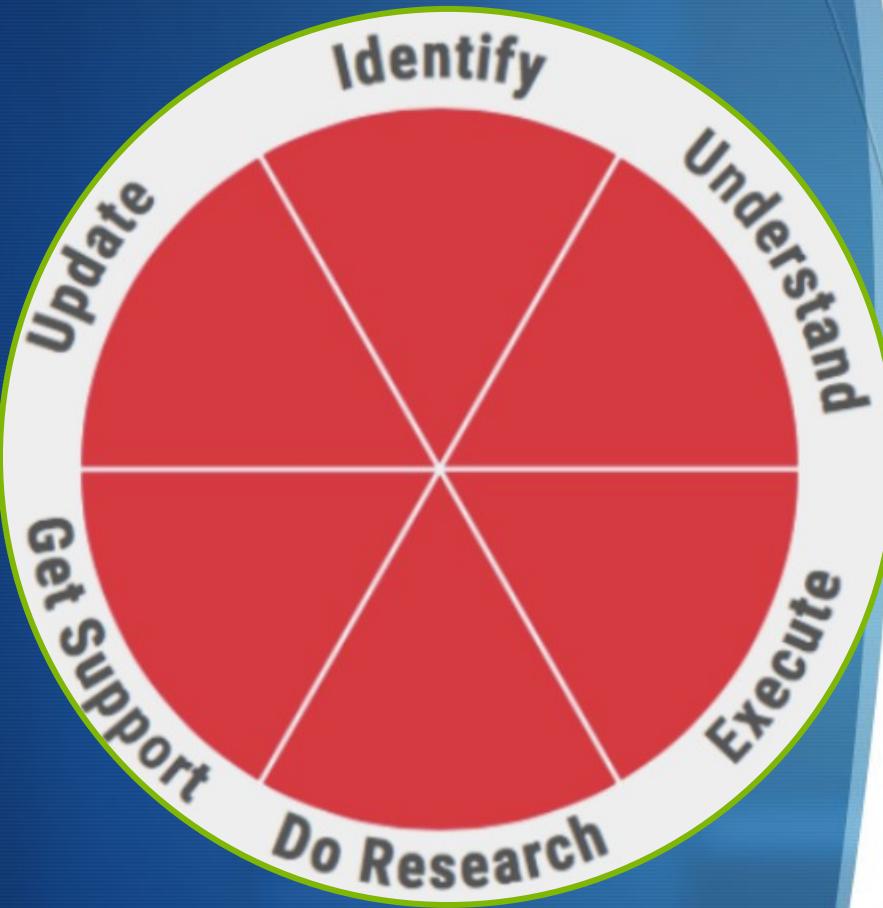
## ★ **Software registry**

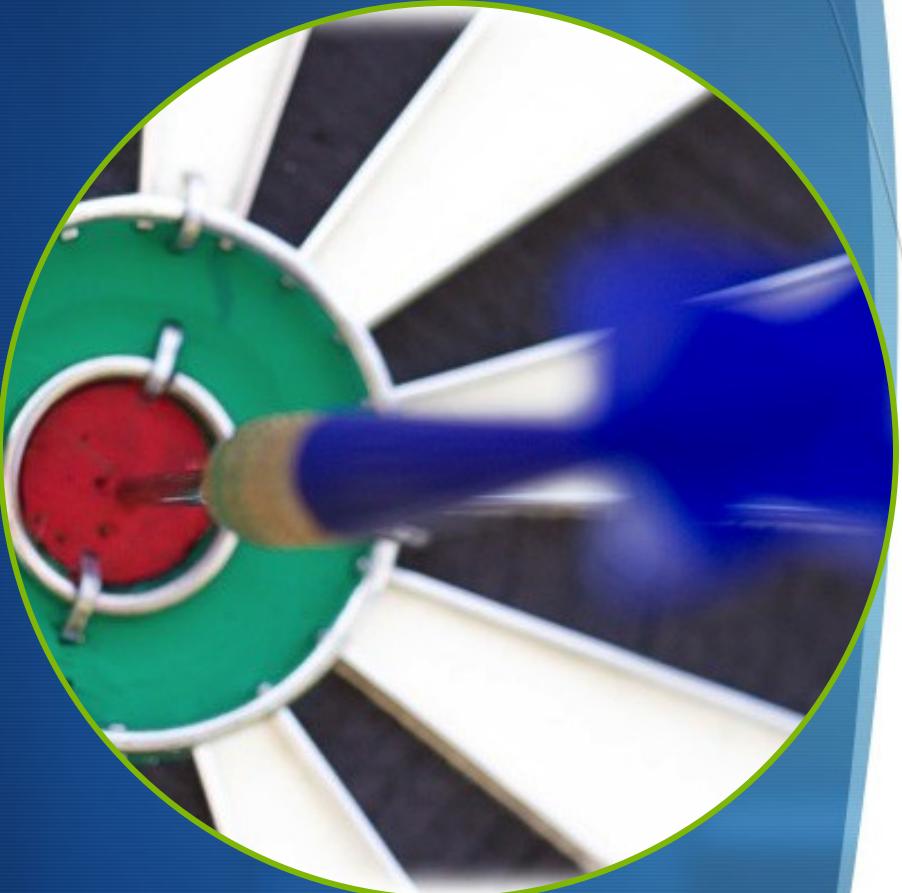
- ★ Capture metadata
  - ★ Useful structured information about the code



# Software Metadata

- ★ Describe characteristics of the software that others can understand, discover (find), and compare software
- ★ Six major categories of software metadata
  - ★ Developed as part of the OntoSoft project
  - ★ <http://www.ontosoft.org/software>





# Goals of this Section

1. Understand what needs to be documented about software to promote reuse
2. Understand how to use a software registry to specify that metadata

# Describing Software with OntoSoft

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

The screenshot shows the 'LOCATE' step of the software identification process. A circular diagram on the left is divided into six segments: 'Identify' (top), 'Understand' (top-right), 'Execute' (right), 'Do Research' (bottom-right), 'Get Support' (bottom-left), and 'Update' (left). Below the diagram, a blue bar says 'Locate unique description'. On the right, there are two tabs: 'Important' (selected) and 'Optional'. Under 'Important', there is a question 'What is the software called ?' followed by the answer 'PIHM'. Under 'Optional', there is a question 'What is a short description for this software ?' followed by a detailed description of PIHM's features and architecture.

Automatic  
crawlers import  
metadata from  
code repositories  
(eg GitHub)

Questions for 6  
top categories,  
some  
“important” and  
some “optional”

The screenshot shows the 'Software Repository' page. It features a search bar at the top right and a 'PUBLISH YOUR SOFTWARE' button. Below is a 'Software List' table with columns for Name, Compare, Edit, and Delete. The table contains entries for DrEICH algorithm, PIHM, PIHMgis, and TauDEM. To the right is a 'Filter Software List' sidebar with dropdowns for Author, Keywords (set to Hydrological model OR Hydrology), and Language (set to C++).

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

# Comparing Alternatives with OntoSoft



Software

Community

Training

## Compare Software

DrEICH algorithm, PIHM, PIHmgis, TauDEM, WBMsed

Select software and features, get a comparison table

PIHM	PIHmgis	DrEICH	TauDEM	WBMsed
<b>What are domain specific keywords for this software ? (eg: hydrology, climate)</b>				
Geomorphology, Hydrological, Bedrock channel ero-	Basins, Continental	Basins, GIS	Hydrologically corrected DEM, Watershed	Sediment flux, Global model, Hydrological model
<b>What Operating Systems can the software run on ?</b>				
Unix Linux	Unix Windows Linux Mac OS	Unix Windows Linux Mac OS	Unix Windows Linux Mac OS	Unix Linux
<b>Is there any test data available for the software ?</b>				
<b>Test Data Location:</b> <a href="http://onlinelibrary.wiley.com/doi/10.1002/2013WR015167/full">http://onlinelibrary.wiley.com/doi/10.1002/2013WR015167/full</a>	<b>Test Data Location:</b> <a href="http://source-forg.net/projects/pihmmodel/">http://source-forg.net/projects/pihmmodel/</a>		<b>Test Data Location:</b> <a href="http://csdms.colorado.edu/wiki/Model:TauDEM#Testing">http://csdms.colorado.edu/wiki/Model:TauDEM#Testing</a>	<b>Test Data Location:</b> <a href="http://csdms.colorado.edu/wiki/Mod-el:WBMsed#Testing">http://csdms.colorado.edu/wiki/Mod-el:WBMsed#Testing</a>
<b>Test Data Description:</b> Two test DEMs are included in the repository,	<b>Test Data Description:</b> Upper Juniata River 875 km^2: see: <a href="http://source-forg.net/projects/pihmmodel/">http://source-forg.net/projects/pihmmodel/</a>		<b>Test Data Description:</b> The Logan River DEM is a small test dataset useful	<b>Test Data Description:</b> Extensive input dataset is available on the CSDMS

# Publishing Software Metadata with OntoSoft

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

OntoSoft

☰

**PIHM**  
[Christopher Duffy]

HTML   RDF/XML   JSON

RATE

**Identify**

**Locate** - Unique description

What is the software called ?

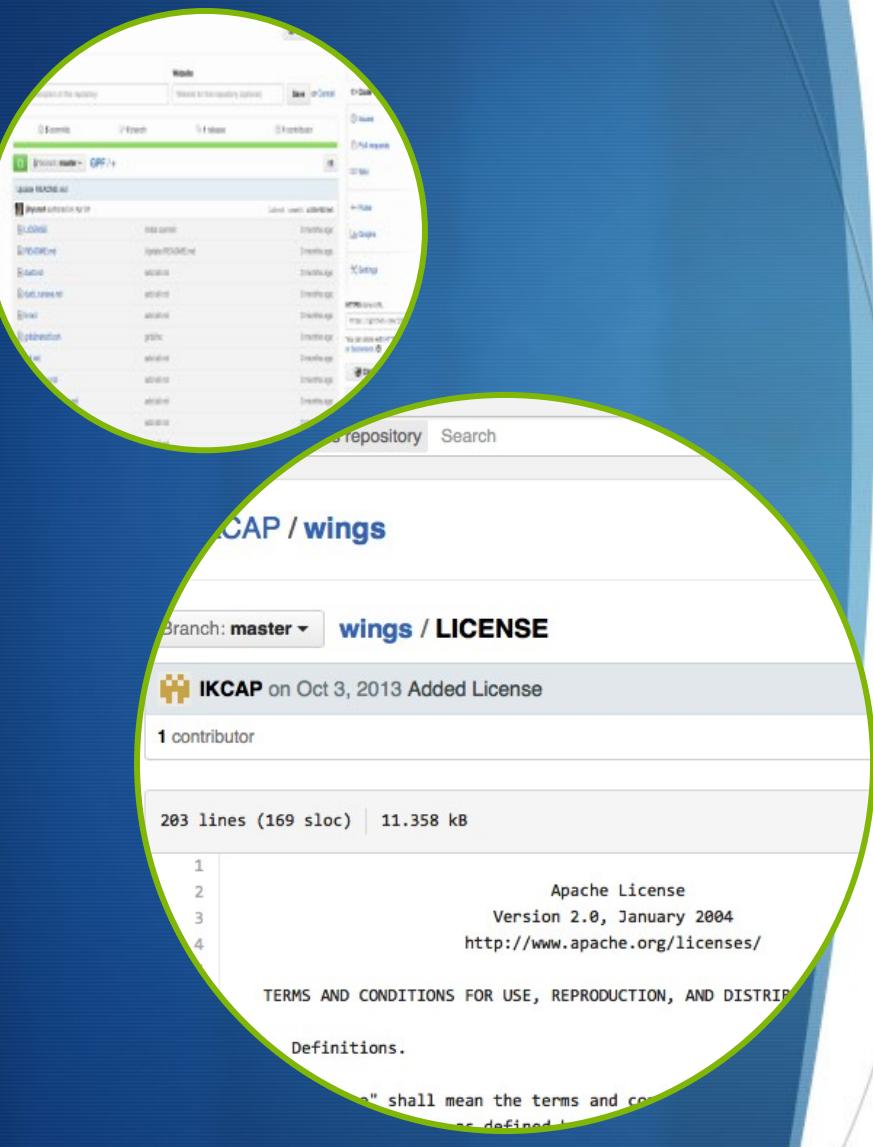
- PIHM

What is a short description for this software ?

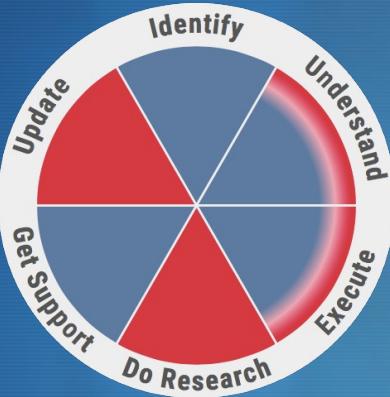
- PIHM is a multiprocess, multi-scale hydrologic model where the major hydrological processes are fully coupled using the semi-discrete finite volume method. PIHM is a physical model for surface and

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

# Documenting Software through Metadata: Simplest Approach



1. Describe as much metadata as you can in your software site
  1. Document the basic metadata discussed earlier
  2. If you use a code repository, there is some basic structure you can follow



What website for the software ?  
[www.pihm.psu.edu/pihm\\_home.html](http://www.pihm.psu.edu/pihm_home.html)

[AL] What is the DOI or any other unique identifier for this software (or software version)?

## Understand

### Trust - Quality and ratings

Who created this software? (Project, Organization, Person, Initiative, etc.)

Christopher Duffy

Are there any additional contributors of note for this software ?

Akash Kumar

Bhatt

Are there any features of this software are worth highlighting ?

Is there anyone else who contributed to this software if not the author ?

# Ideal Approach

1. Use a software registry
  - <http://www.ontosoft.org/portal>, csdms.colorado.edu, etc.
  - Guides through questions to provide metadata
2. Save the metadata as HTML, XML,...
3. Post the metadata on your code site

# g Provenance and Methods

## OntoSoft Training

### Part 5

<http://dx.doi.org/10.5281/zenodo.15920>

<http://www.scientificpaperofthefuture.org>



CC-BY  
Attribution



# Methods Described in Text Are Incomplete and Ambiguous

- ★ Analysis of 18 quantitative papers published in Nature Genetics in the past two years found that reproducibility was not achievable even in principle in 10 cases, even when datasets are published [Ioannidis et al 09]
- ★ “Data processing, however, is often not described well enough to allow for exact reproduction of the results, leading to exercises in ‘**forensic bioinformatics**’ where aspects of raw data and reported results are used to infer what methods must have been employed.” [Baggerly and Coombes 09]
- ★ “**Ambiguity** in program descriptions leads to the possibility, if not the certainty, that a given natural language description can be converted into computer code in various ways, each of which may lead to different numerical outcomes.” [Ince et al 2012]

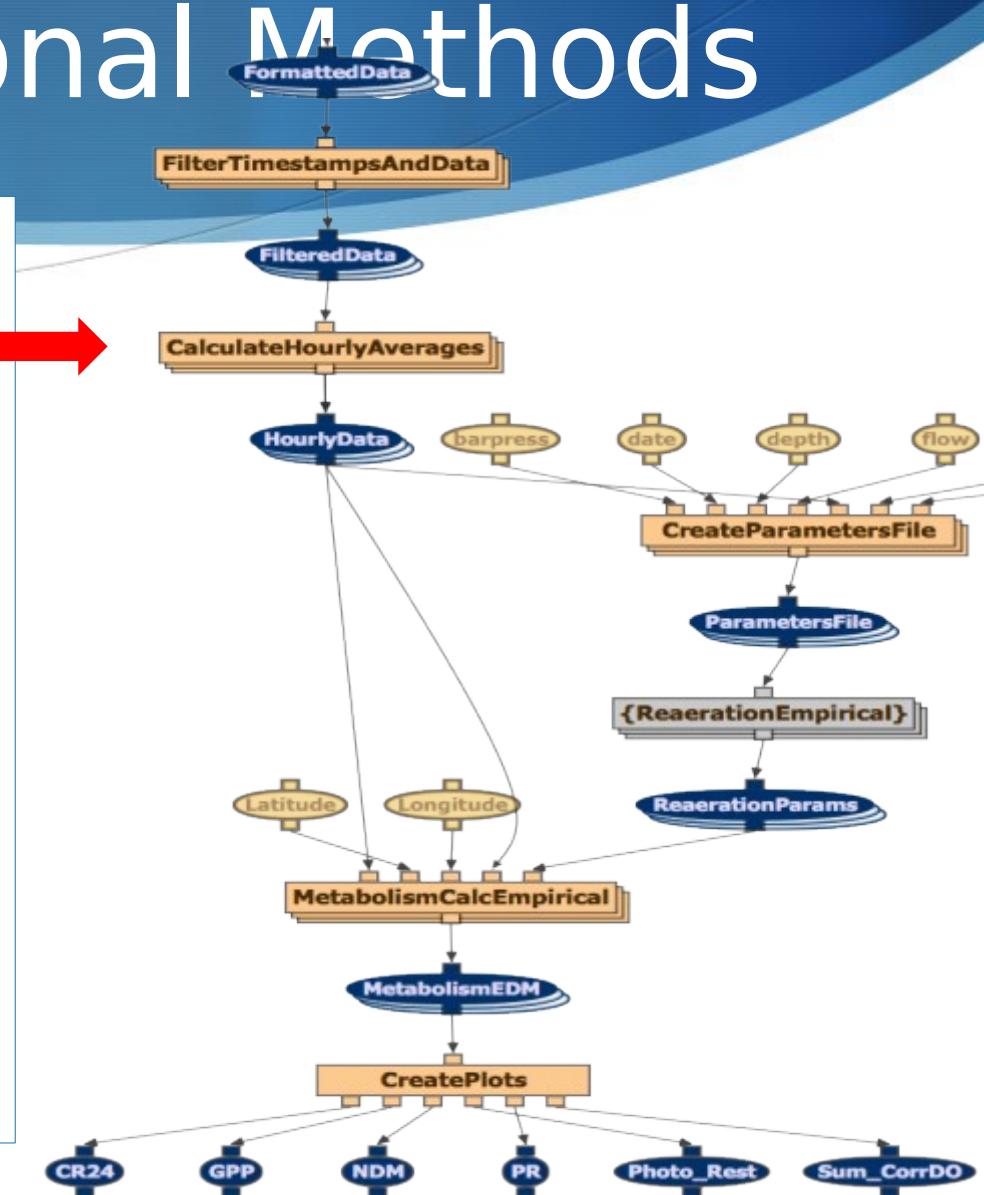


# Goals of this Section

1. Understand what are methods and provenance is in a scientific article
2. Understand how to document methods and provenance properly in an article

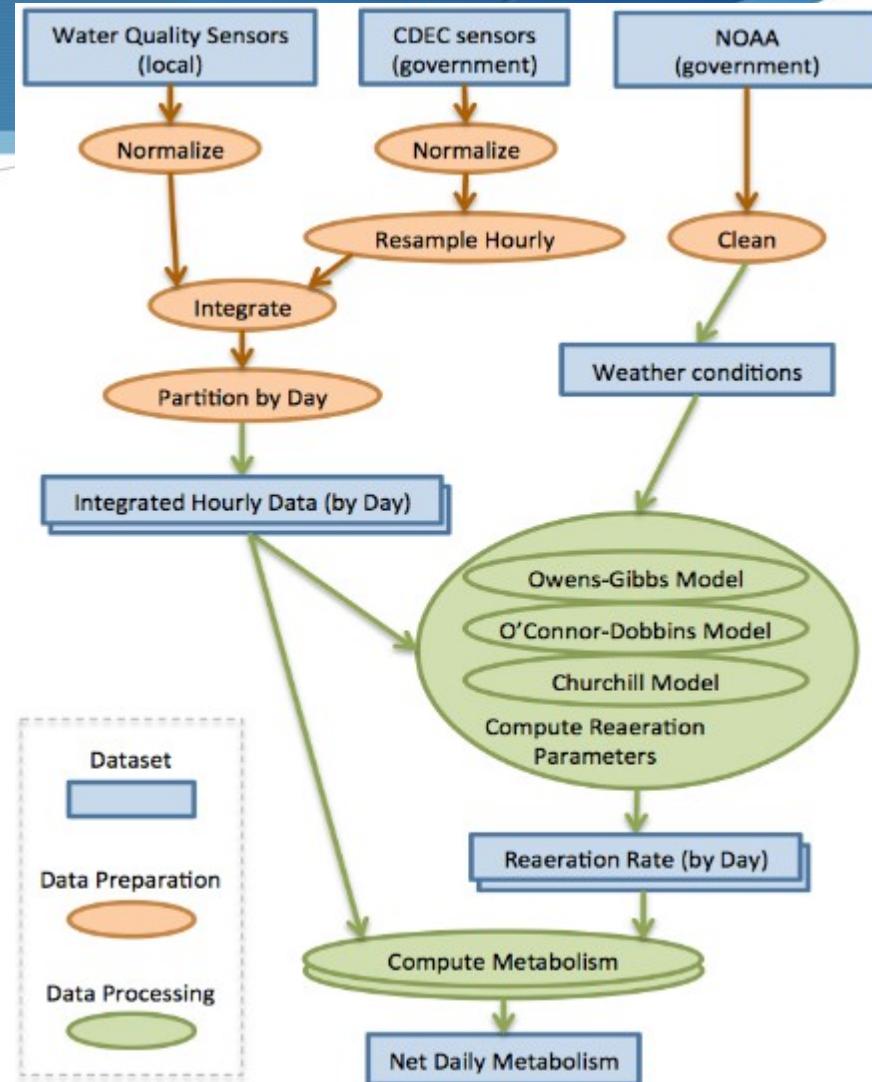
# Workflows as Representations of Computational Methods

- ★ Computational workflow
  - ★ Eg, water metabolism
- ★ Workflows can include manual steps
  - ★ Eg, creating a figure, cleaning data
- ★ Workflows may access web services
  - ★ Eg, access databases in biology

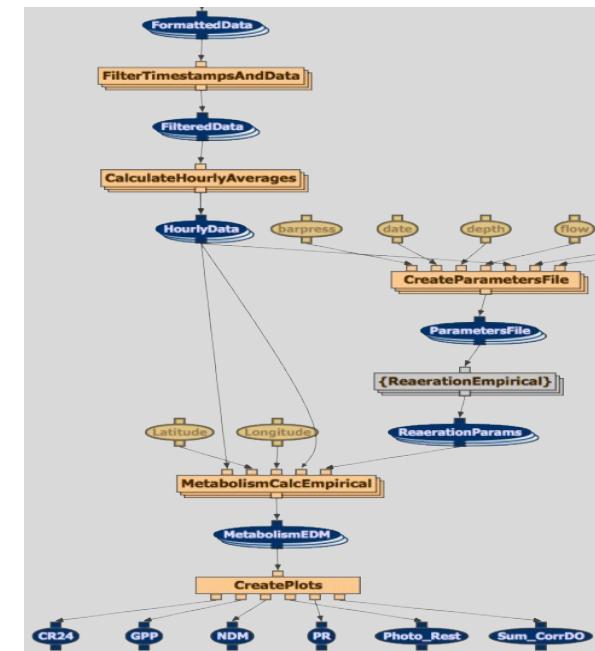
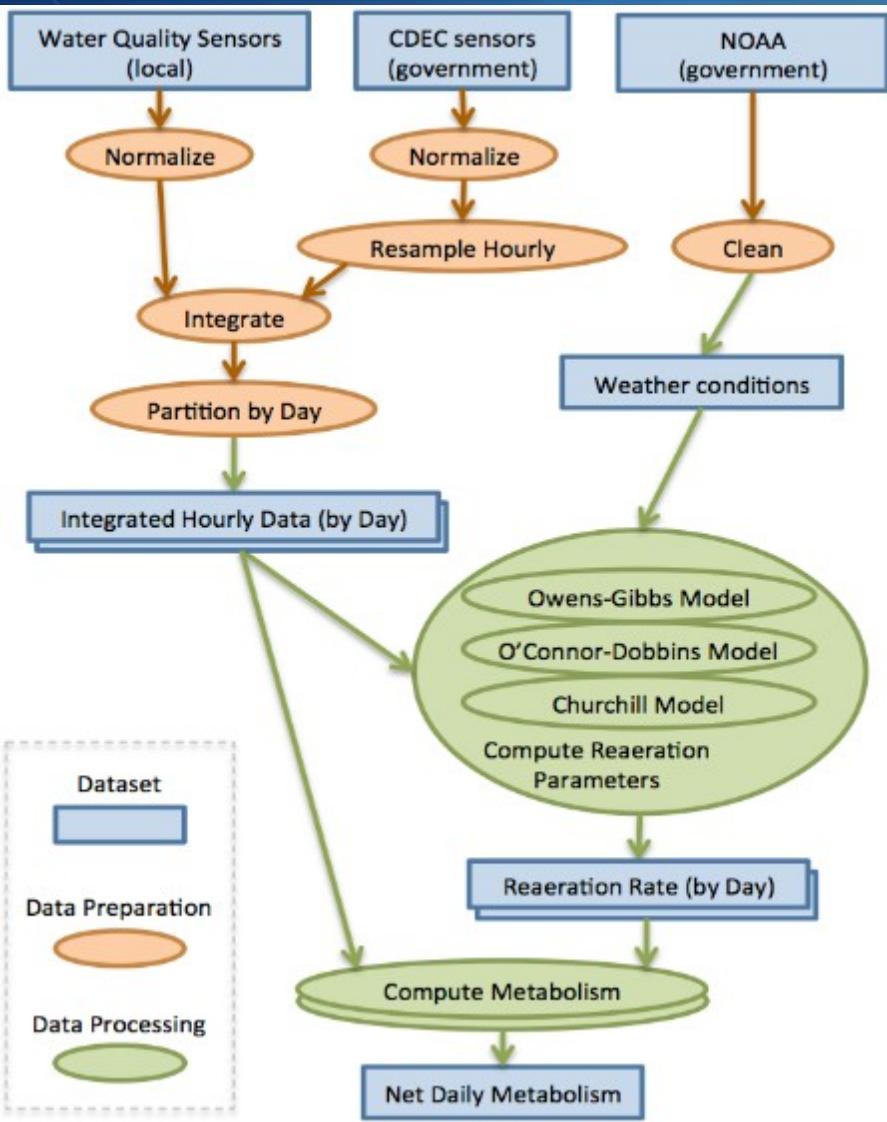


# Developing Workflows: How to Sketch a Workflow

1. Compile the command line invocation to all your codes
  - ★ Input data, parameters, configuration files
  - ★ Include data preparation codes
2. Consider how the data flows from code to code
3. Starting with the input data, work your way to the results
4. If any steps were done with manual intervention, indicate that
5. Create subworkflows if it gets large



# From a Workflow Sketch to a Formal Workflow



# Workflow Systems

- ★ Capture method as a workflow
- ★ Workflow can be easily shared and reused
- ★ Other benefits
  - ★ Workflow validation
  - ★ Scalable computations
  - ★ Comprehensive software libraries
- ★ Many workflow systems
  - ★ Each has different capabilities



# Electronic Notebooks

## IP[y]: Notebook

Sweave = R · LATEX

## CDF

### Computable Document Format

Documents come alive with the power of computation



IP[y]: Notebook    spectrogram    Last saved: Mar 07 11:14 PM

File Edit View Insert Cell Kernel Help

Simple spectral analysis

An illustration of the [Discrete Fourier Transform](#)

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} kn} \quad k = 0, \dots, N - 1$$

using windowing, to reveal the frequency content of a sound signal.

We begin by loading a datafile using SciPy's audio file support:

```
In [1]: from scipy.io import wavfile  
rate, x = wavfile.read('test_mono.wav')
```

And we can easily view its spectral structure using matplotlib's builtin `specgram` routine:

```
In [2]: fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(12, 4))  
ax1.plot(x); ax1.set_title('Raw audio signal')  
ax2.specgram(x); ax2.set_title('Spectrogram');
```

1



3



2



# What is Provenance

Provenance covers:

1. Processes
2. Documents (“resources”)
3. Entities

[http://en.wikipedia.org/wiki/Certificate\\_of\\_origin#mediaviewer/  
File:Coal\\_from\\_the\\_Titanic.jpg](http://en.wikipedia.org/wiki/Certificate_of_origin#mediaviewer/File:Coal_from_the_Titanic.jpg)

[http://commons.wikimedia.org/wiki/File:The\\_seal\\_of\\_National\\_Taiwan\\_University.png](http://commons.wikimedia.org/wiki/File:The_seal_of_National_Taiwan_University.png)

[https://www.flickr.com/photos/altorschwede/2303630740/ \(CC BY ND 2.0\)](https://www.flickr.com/photos/altorschwede/2303630740/)

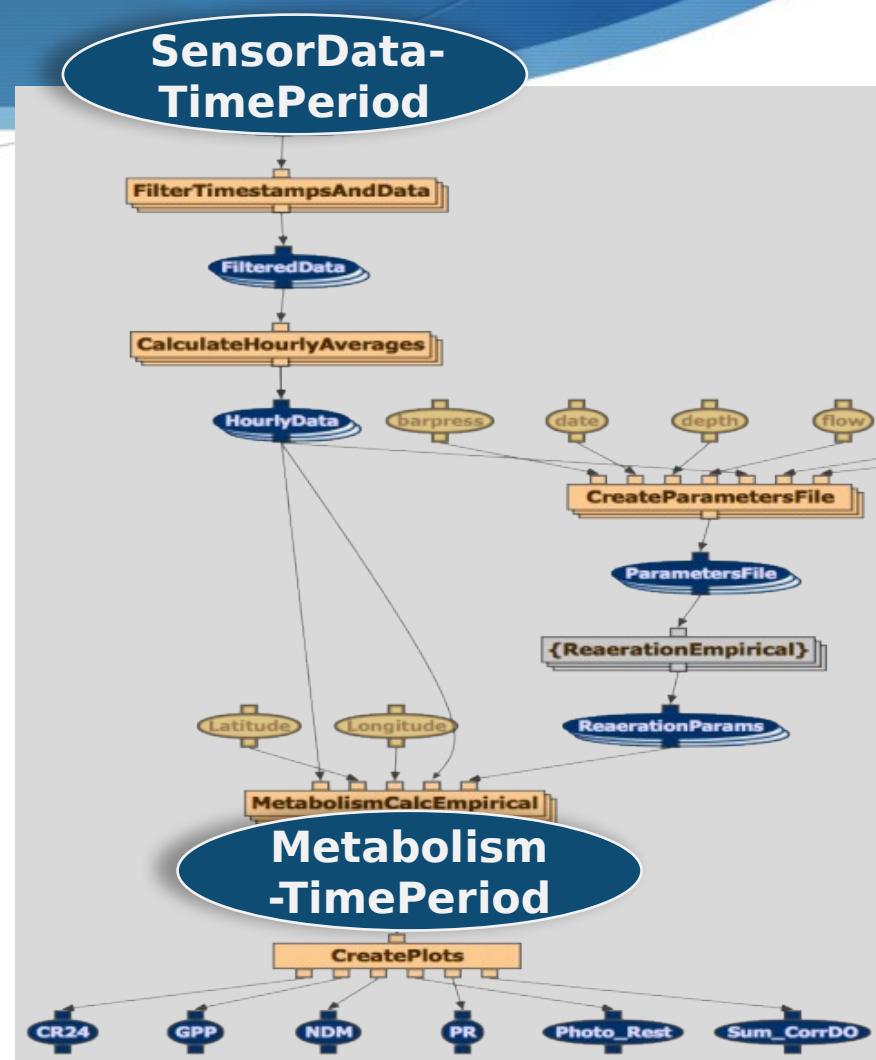
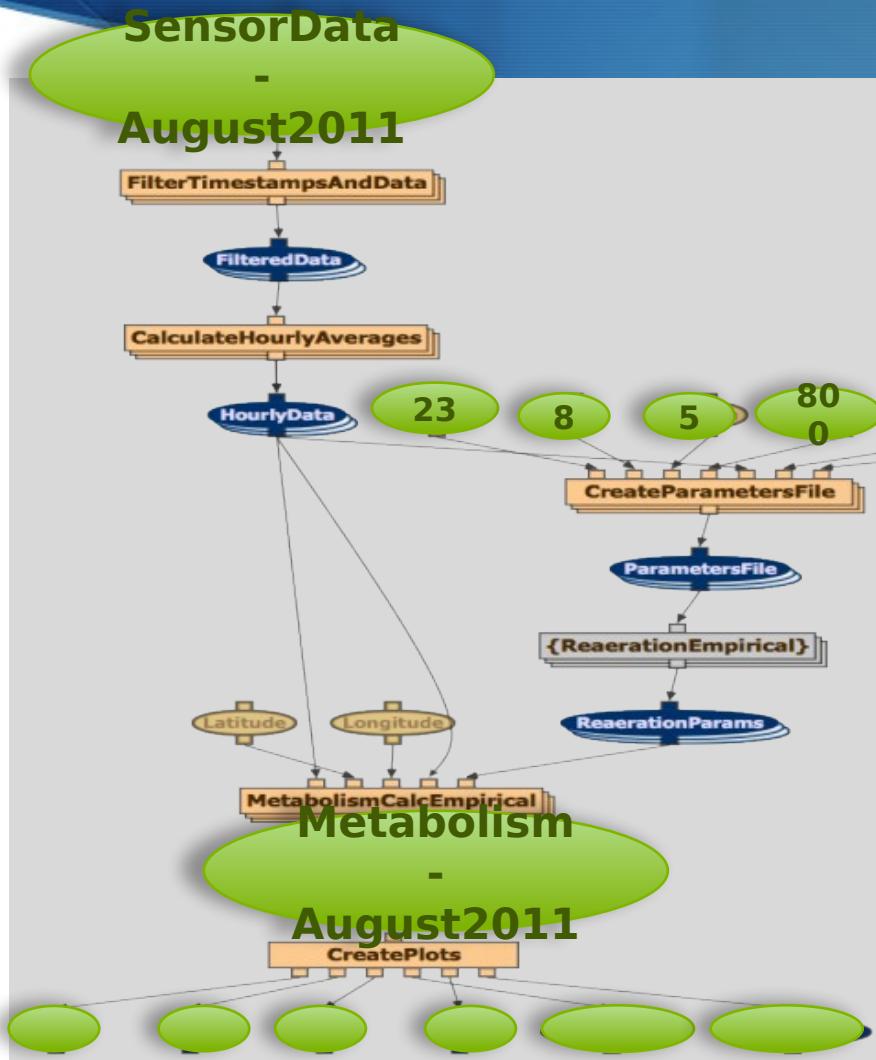
# A Working Definition of Provenance

Provenance of a resource is **a record** that describes entities and processes involved in producing and delivering or otherwise influencing that resource.

Provenance provides a critical foundation for assessing authenticity, enabling trust, and allowing **reproducibility**.

- ★ Provenance results from **past** actions
  
- ★ Provenance can be seen as **metadata**, but not all metadata is provenance

# Describing Execution (Provenance) vs General Method (Workflow)

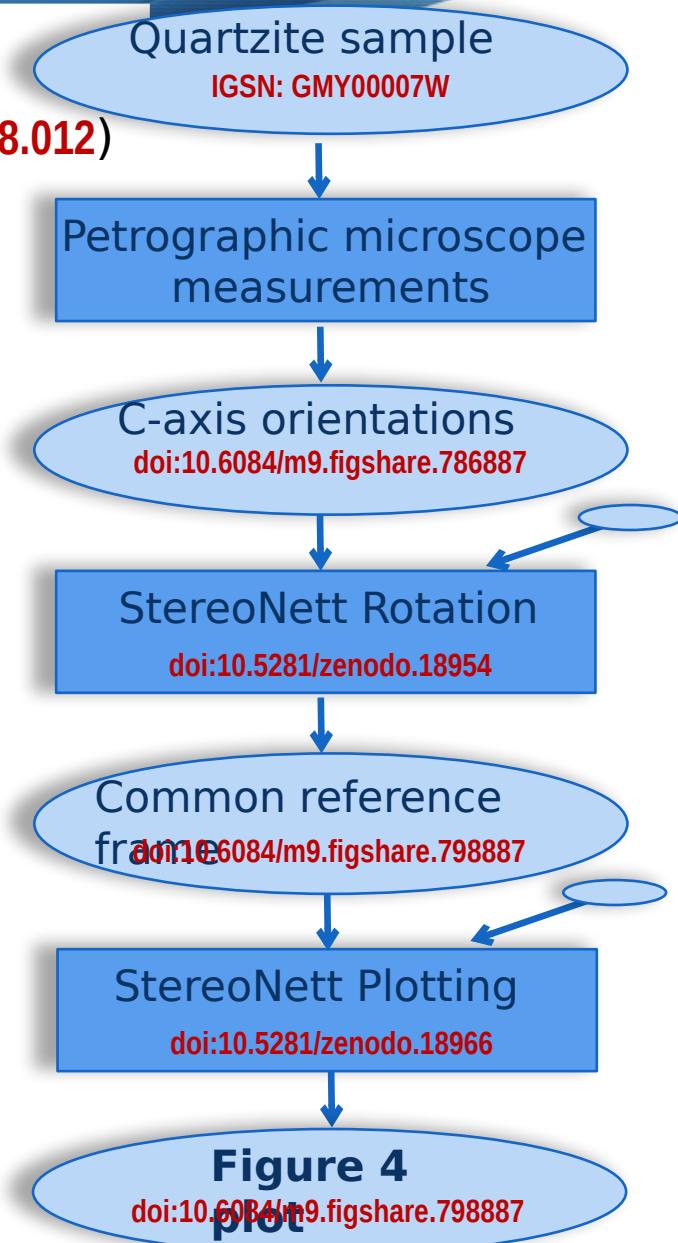


# Example: Text and

## Understanding kinematic data from the Moine thrust zone ([doi:10.1016/j.ess.2009.08.012](https://doi.org/10.1016/j.ess.2009.08.012))

Jade Silverstein ([orcid.org/0000-0001-8455-8431](https://orcid.org/0000-0001-8455-8431))

[...] We took a quartzite sample (**IGSN:** **GMY00007W**) from the Stack of Glencoul in the Moine thrust, and cut 3 thin sections. We measured c-axis orientations ([doi:10.6084/m9.figshare.786887](https://doi.org/10.6084/m9.figshare.786887)) using a petrographic microscope. We rotated to a common reference frame ([doi:10.6084/m9.figshare.798887](https://doi.org/10.6084/m9.figshare.798887)) using Duyster's StereoNett program ([doi:10.5281/zenodo.18954](https://doi.org/10.5281/zenodo.18954)). We plotted the data on lower hemisphere, equal area projections ([doi:10.6084/m9.figshare.798887](https://doi.org/10.6084/m9.figshare.798887)) using Duyster's StereoNett program ([doi:10.5281/zenodo.18966](https://doi.org/10.5281/zenodo.18966)).





# Goals of this Section

1. Understand what are methods and provenance is in a scientific article
2. Understand how to document methods and provenance properly in an article

# 1

by a scoring function to determine the statistical significance of the statistical model derived from the data.

Software was used to compare the pharmacology models (a total of 2,195 drugs, in an all-against-all manner) defined by the bound ligand, the receptor was scanned in order to generate a representation of the receptor.

# 3

```
storage/users/admin/Water/code/library/CreateParametersFileNode_9  
-----  
/usr/share/tomcat6/storage/users/admin/Water/code/library/CreateParametersFileNode_9  
/usr/share/tomcat6/storage/users/admin/Water/data/AvgHourly_SMN_2010-03-03Z  
  
ReaerationCMNode  
-----  
/usr/share/tomcat6/storage/users/admin/Water/code/library/ReaerationCM/run -o1  
/usr/share/tomcat6/storage/users/admin/Water/data/Params_SMN_2010-03-03Z  
/usr/share/tomcat6/storage/users/admin/Water/code/library/ReaerationCM/run -o1  
/usr/share/tomcat6/storage/users/admin/Water/data/Params_SMN_2010-03-03Z  
  
CreateParametersFileNode  
-----  
/usr/share/tomcat6/storage/users/admin/Water/code/library/CreateParametersFileNode_5  
/usr/share/tomcat6/storage/users/admin/Water/data/AvgHourly_SMN_2010-03-03Z  
  
CreateParametersFileNode_5  
-----  
/share/tomcat6/storage/users/admin/Water/code/library/CreateParametersFileNode_5  
/share/tomcat6/storage/users/admin/Water/data/AvgHourly_SMN_2010-03-03Z  
  
AveragesNode_6  
-----  
/share/tomcat6/storage/users/admin/Water/code/library/AveragesNode_6  
/share/tomcat6/storage/users/admin/Water/data/AvgHourly_SMN_2010-03-03Z
```



## Documenting Provenance and Methods:

# Simplest Approach

## 1. Describe the workflow in text

- Data + software + workflow
- Specify unique identifiers for data and software, versions, credit all sources

## 2. Develop a workflow sketch

- Capture high-level dataflow across components

## 3. For provenance, include a summary or an execution

# Ideal Approach

1

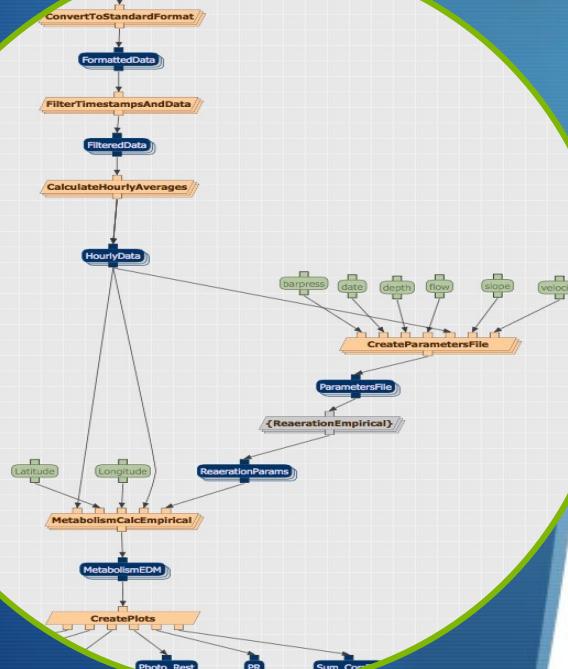
by a scoring function to determine the statistical significance of the statistical model derived from the data.

Software was used to compare the pharmacology models (a total of 2,195 drugs, in an all-against-all manner) defined by the bound ligand, the receptor was scanned in order to generate a representation of the

2



3



1. Describe the workflow in text
  - Data + software + workflow
  - Specify unique identifiers for data and software, versions, credit all sources
2. Develop a workflow sketch
  - Capture high-level dataflow across components
3. Specify the formal workflow using a workflow system, electronic notebook, etc.
  - Command lines + parameter values
  - Dataflow across components
4. Include the provenance record
  - If generating it automatically, preferably using a standard (e.g., PROV)
5. Publish the workflow and provenance record in a publicly accessible repository (eg figshare, myExperiment, etc)
6. Get a unique persistent identifier for the workflow, the provenance, or both

# Documenting Provenance and Methods:

## How to show provenance and workflow in the article

- ★ Describe the workflow in text
  - ★ In the “Methods” section
- ★ Include your workflow sketch
  - ★ As a figure in the article
- ★ Include your provenance summary or trace
- ★ If available as formal workflow and provenance record, cite them in the paper (use a format analogous to data and software citation)

# The Scientific Paper of the Future: An Author Checklist

## OntoSoft Training

### Part 6

<http://dx.doi.org/10.5281/zenodo.15920>



<http://www.scientificpaperofthefuture.org>

CC-BY  
Attribution



# Review of Best Practices: A GPF Author Checklist

1

Data accessibility

2

Data documentation

3

Software accessibility

4

Software  
documentation

5

Provenance  
documentation

6

Methods  
documentation

7

Authors identification

# What to Show in a GPF

## Data Citation Format

Cite this:

Garijo, Daniel; Xie, Lei; Zhang, Yinliang; Gil, Yolanda; Xie, Li; Kinnings, Sarah;  
Bourne, Phil (2013) Highly connected drug file figshare.  
<http://dx.doi.org/10.6084/m9.figshare.776887>  
Retrieved 11/05, Feb 20, 2015 (GMT)

Authors

Date of publication

Time of retrieval

- ★ Cite each of your datasets like you would cite another paper
- ★ Citation includes publication date, date of retrieval, repository, and persistent identifier
- ★ If there is a data paper, cite it

Permanent unique identifier

Name

Repository

# What to Show in a GPF

The US  
Long Term Ecological Research  
Network

LTER Identifier:  
knb-lter-ntl.279.1

Abstract:  
These data were collected by the Wisconsin Department of Natural Resources (WDNR) from 1987-1998. Most of these data (1987-1993) precede 1995, the year that the University of Wisconsin A NTL-LTER program A took over sampling of the Yahara Lakes. However, WDNR data collected from 1997-1998 A (unrelated to LTER sampling) is also included. In 1987 a joint project by the WDNR and the University of Wisconsin-Madison, Center for Limnology (CFL) was initiated on Lake Mendota. The project involved biomimicry ...

Owners/Creators:  
Lathrop

Metadata:  
Select [here](#) for full metadata

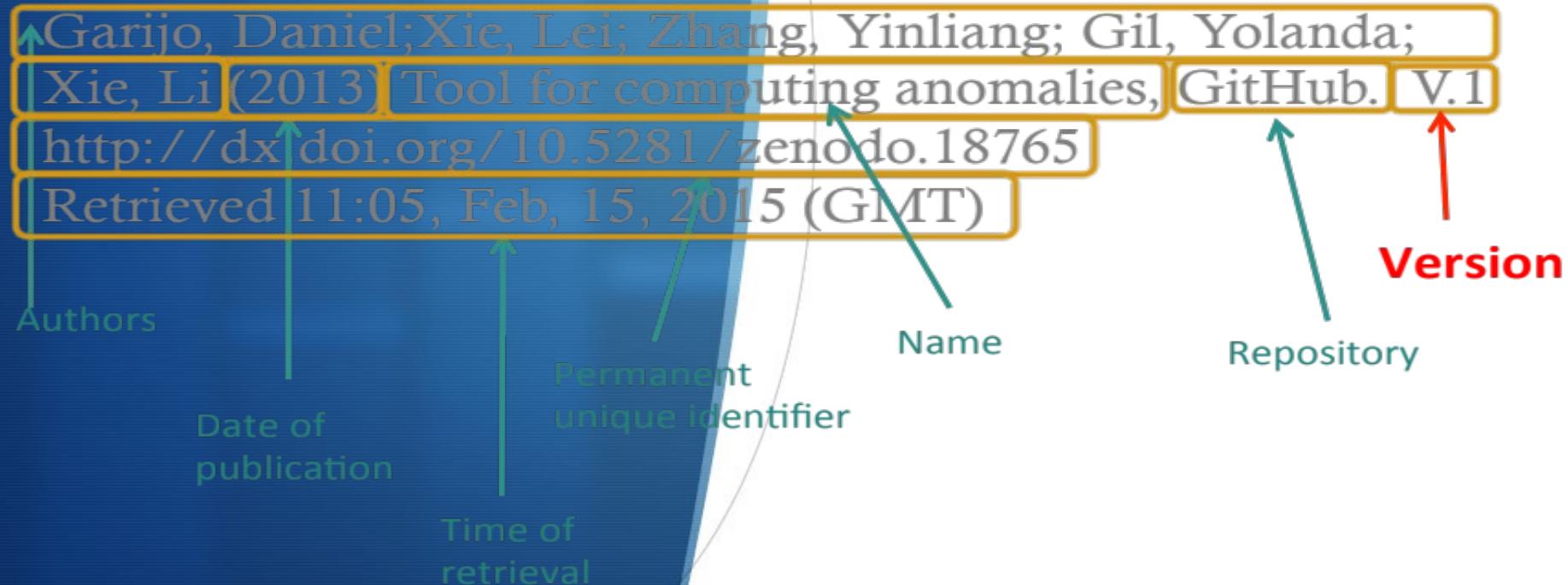
Data File(s):

- [wdnr\\_fyke\\_minifyke\\_seine\\_lengths\\_weights.csv](#)
- [wdnr\\_boomshock\\_lengths\\_weights.csv](#)
- [wdnr\\_gillnet\\_lengths\\_weights\\_93.csv](#)
- [wdnr\\_walleye\\_age\\_lengths\\_weights\\_87.csv](#)
- [wdnr\\_creel\\_survey\\_lengths\\_weights.csv](#)
- [wdnr\\_creel\\_survey\\_angler\\_counts.csv](#)

- ★ Mention that the persistent identifier for your data has pointers to its metadata and includes a detailed description of the data
- ★ Optionally, include the metadata also as supplemental material
- ★ If there is a data paper, cite it

# What to Show in a GPF

## Software Citation Format



- ★ Cite each piece of software that you use (preparation, analysis, visualization) like you would cite another paper
- ★ Citation similar to data but includes software version
- ★ If there is a software paper, cite it

# What to Show in a GPF

- ★ Mention that the persistent identifier location for your software points to its metadata
- ★ Optionally, include the software metadata as supplemental material
- ★ If there is a software paper, cite it

## PIHM [Christopher Duffy]

### Identify

#### Locate - Unique description

What is the software called ?

- PIHM

What is a short description for this software ?

- PIHM is a multiprocess, multi-scale hydrologic model where the major hydrological processes are fully coupled using the semi-discrete finite volume method. PIHM is a physical model for surface and groundwater, tightly-coupled to a GIS interface. PIHMgis which is open source, platform independent and extensible. The tight coupling between GIS and the model is achieved by developing a shared data-model and hydrologic-model data structure.

Initial metadata was retrieved from <http://csdms.colorado.edu/wiki/Model:PIHM>

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

- Hydrology
- Basins
- Continental

Is there a project website for the software ?

- [http://www.pihm.psu.edu/pihm\\_home.html](http://www.pihm.psu.edu/pihm_home.html)

### Understand

#### Trust - Quality and ratings

Who created this software? (Project, Organization, Person, Initiative, etc.)

- Christopher Duffy

Are there any additional contributors of note for this software ?

- Mukesh Kumar
- Gopal Bhatt

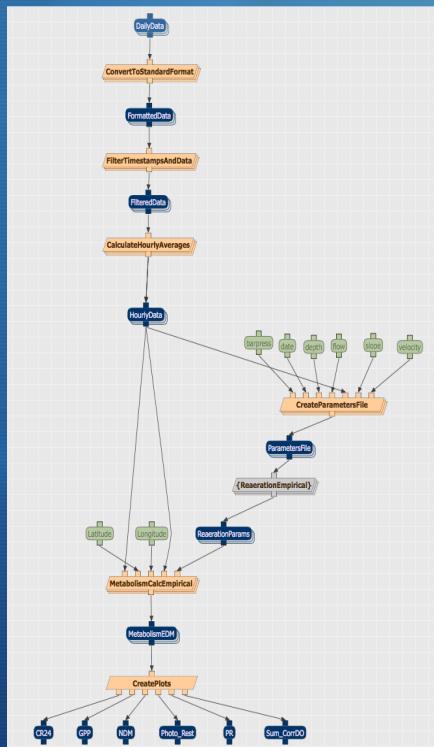
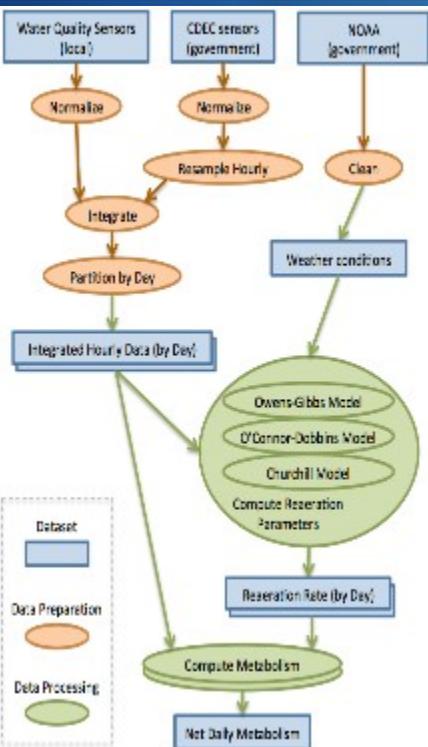
5

Provenance documentation

6

Methods documentation

# what to Show in a GPF



- ★ Describe workflow in text and provide a workflow sketch

- ★ Optionally, provide the formal workflow or lab notebook, use a persistent identifier, and cite it
- ★ Include a summary of the execution traces as supplementary material, or use a persistent identifier and cite it

- ★ Optionally, include instead the provenance using a standard PROV

**# Entities**

```

ex:testData1 a prov:Entity .
ex:model1 a prov:Entity .
ex:classification1 a prov:Entity .

```

**# Activities**

```

ex:Classifier1 a prov:Activity .

```

**# Usage and Generation relations between entities and activities**

```

ex:Classifier1 prov:used ex:testData1 ;
ex:used ex:model1 .

ex:Classification1 prov:wasGeneratedBy ex:Classifier1 .

```

# What to Show in a GPF



- ★ Authors have a persistent unique identifier
  - ★ Use [www.orcid.org](http://www.orcid.org)
  - ★ Instructions are on the AGU ESS journal GPF special issue web site

ORCID

# A GPF Author Checklist

1

Data accessibility

2

Data documentation

3

Software accessibility

4

Software documentation

5

Provenance documentation

6

Methods documentation

7

Authors identification

- ★ **For datasets**, the paper should include one or more citations, specifying the authors, the site where they are described and can be accessed, the repository, and the license.
- ★ **For software**, the paper should include one or more citations, specifying the authors, the site where it is described and can be accessed, the repository, and the license.
- ★ **For provenance and workflow**, the paper should include figures and traces, and if available the citations mentioning the authors, site to access them, the repository, and the license.

# Today: To Write a Scientific Paper of the Future and also to

- 1. Get credit** for all your research products
  - ★ Citations for software, data, samples, ...
- 2. Increase citations** of your papers
- 3. Write impressive Data Management Plans**
- 4. Extend your CV** with data and software sections
- 5. Reproduce** your work from years ago
- 6. Comply with new funder and journal requirements**

# Incorporate GPF Best Practices Into Your Work



- Easier to track research products, report to funders, get credit, etc.
- Making a paper into a GPF is then very straightforward

# Acknowledgments



EarthCube



- ★ The Scientific Paper of the Future training materials were developed and edited by Yolanda Gil (USC), based on the OntoSoft Geoscience Paper of the Future (GPF) training materials with contributions from the OntoSoft team including Chris Duffy (PSU), Chris Mattmann (JPL), Scott Peckham (CU), Ji-Hyun Oh (USC), Varun Ratnakar (USC), Erin Robinson (ESIP)
- ★ The OntoSoft training materials were significantly improved through input from GPF pioneers Cedric David (JPL), Ibrahim Demir (UI), Bakinam Essawy (UV), Robinson W. Fulweiler (BU), Jon Goodall (UV), Leif Karlstrom (UO), Kyo Lee (JPL), Heath Mills (UH), Suzanne Pierce (UT), Allen Pope (CU), Mimi Tzeng (DISL), Karan Venayagamoorthy (CSU), Sandra Villamizar (UC), and Xuan Yu (UD)
- ★ Thank you to Ruth Duerr (NSIDC), James Howison (UT), Matt Jones (UCSB), Lisa Kempler (Matworks), Kerstin Lehnert (LDEO), Matt Meyernick (NCAR), and Greg Wilson (Software Carpentry) for feedback on best practices
- ★ Thank you also to the many scientists and colleagues that have taken the training and asked hard questions
- ★ We are grateful for the support of the National Science Foundation and

# For More Information

<http://www.scientificpaperofthefuture.org>



<http://dx.doi.org/10.5281/zenodo.159>

