

The Scientific Paper of the Future

<http://www.scientificpaperofthefuture.org>

OntoSoft Training

January 2017

ontosoft@gmail.com

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



Onto
Soft



[CER-1440323]

[CER-1343800]

CC-BY
Attribution



EarthCube

Instructors Today

Daniel Garijo



Yolanda Gil



Gail Clement



**Information Sciences Institute
University of Southern California**

**Research Library
California Institute
of Technology**

Acknowledgments



ICER-1440323
ICER-1343800

- ★ 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 the EarthCube program



ICER-1440323
ICER-1343800

OntoSoft: Software Stewardship for the Geosciences

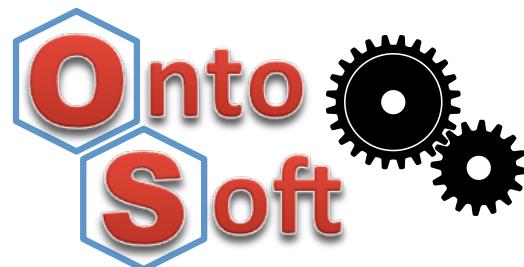


EarthCube



Community

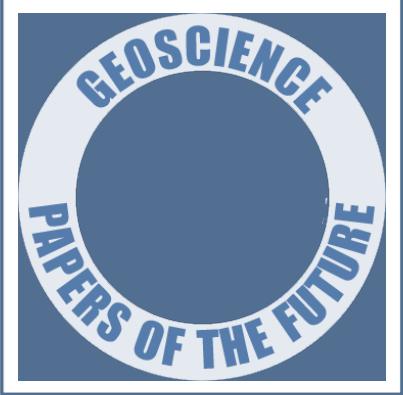
- Recommender system
- Interoperability



Publication

- Structured metadata
- Interactive advice





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



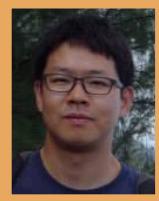
R. W. Fulweiler, Boston U.
Biogeochemistry in marine ecology



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 support



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>



Onto
Soft



[CER-1440323]
[CER-1343800]

CC-BY
Attribution



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

Scientists Are Changing

Open data



Open source



Open publications



Open access



Scientists Are Changing

NATURE METRICS SURVEY 2010

METRICS SURVEY RESULTS

Thinking about all of the possible measures of scientific contribution that are possible, please select your top 5 priorities.

	No. of times chosen	Relative ranking
Publication in high-impact journals	92	2.61
Grants earned	65	1.73
Training and mentoring students and postdocs	63	1.71
No. of citations on published research	58	1.62
No. of publications	53	1.38
Teaching courses	41	1.18
Collaborative work outside of your department/institution	37	0.97
Development of research resources for the scientific community	31	0.89
Invitations to talk at meetings	29	0.80
Collaboration/cooperation within your department/institution	25	0.66
No. of students or postdocs who go on to prestigious jobs	25	0.63

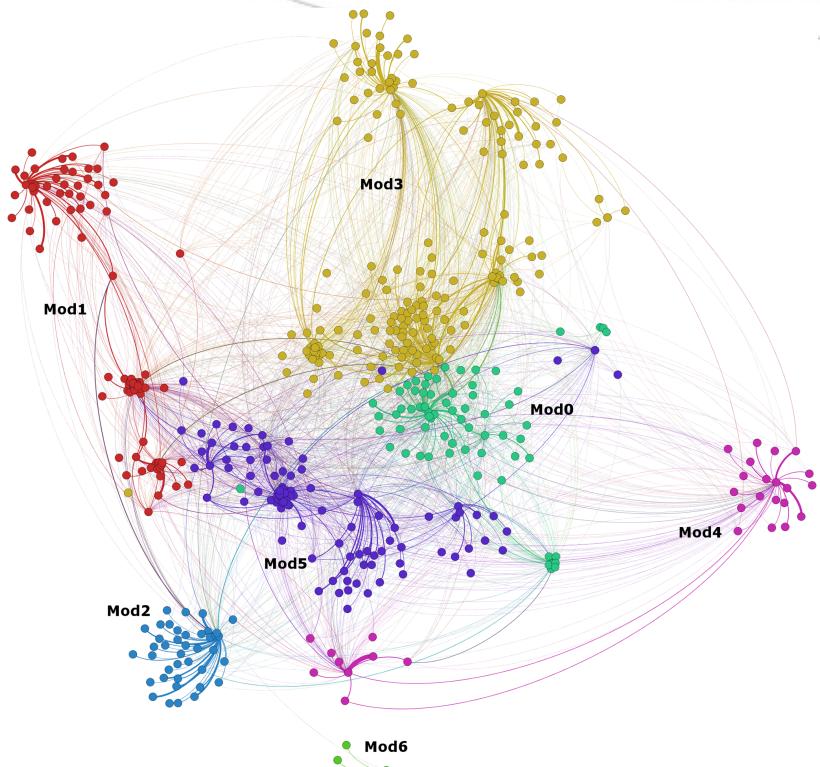
Thinking about all of the possible measures of scientific contribution that are possible, please select your top 5 priorities.

No. of times chosen Relative ranking

Publication in high-impact journals	92	2.61
Grants earned	65	1.73
Training and mentoring students and postdocs	63	1.71
No. of citations on published research	58	1.62
No. of publications	53	1.38
Teaching courses	41	1.18
Collaborative work outside of your department/institution	37	0.97
Development of research resources for the scientific community (e.g. reagents, software, database development)	31	0.89

Departmental/institutional administration	5	0.16
Development of start-up business	5	0.14
Blogging, writing for lay press	4	0.10
Meeting abstracts	3	0.08
Data deposited in public repositories	3	0.08
Participation in departmental meetings	2	0.05

Scientists Are Changing



[Holmberg et al 2014]

<http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0106086>



Altmetric



Crossref



Impactstory



Publishers Are Changing



2 December 2011 | \$10

Data Replication & 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

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 another researcher can reproduce the experiments described, (2) our aim to promote openness, and (3) the fact that PLOS journals can be built upon by future researchers. Therefore, if new software or a new application is developed and used in a paper, 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, if the source code is made available.
- **Documentation for running and installing the software.** For end-user applications, instructions for installing the software are sufficient; for software libraries, instructions for using the application program interface (API) are sufficient.
- **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.

The Public is Changing



eBird



Discovery of Western European R1b1a2 Y Chromosome Variants in 1000 Genomes Project Data: An Online Community Approach

Richard A. Rocca , Gregory Magoon, David F. Reynolds, Thomas Krahn, Vincent O. Tilroe, Peter M. Op den Velde Boots, Andrew J. Grierson

Published: July 24, 2012 • DOI: 10.1371/journal.pone.0041634

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;

Funders Are Changing

NSF'S PUBLIC ACCESS PLAN:

Today's Data, Tomorrow's Discoveries

Increasing Access to the Results of Research Funded by the
National Science Foundation

National Science Foundation

March 18, 2015



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.



★ Data published in a repository

 The US Long Term Ecological Research Network

+/- NTL LTER "WDNR Yahara Lakes Fisheries: Fish Lengths and Weights 1987-1998" - Lathrop

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 Å NTL-LTER program Å took over sampling of the Yahara Lakes. However, WDNR data collected from 1997-1998 Å (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 biomonitoring 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_creel_survey_lengths_weights.csv](#)
- [wdnr_creel_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

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

**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](#).

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

By Paul Goldberg

Biostatistics journals aren't usually the place to find sensational claims about medical treatments. But the most recent issue of the Annals of Applied Statistics is an exception.

Methodology

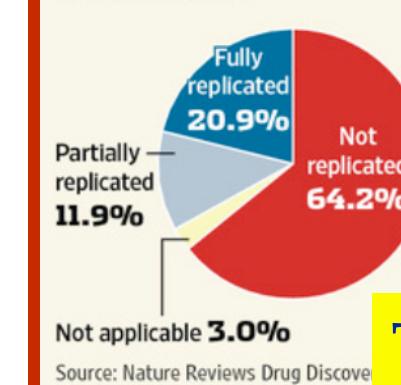
A paper published in the journal claims that patients in two Duke University clinical trials may be harmed by a common biomarker used to predict heart attacks. The paper's lead author, a biostatistician at the University of Michigan, says he relied on biomarker data from the trials to show that the biomarker was unreliable.

The paper has not been peer-reviewed, and Duke University has not yet responded to requests for comment. The journal's editor, Robert Tibshirani, says he has not yet had time to review the paper.

Biostatistics journals aren't usually the place to find sensational claims about medical treatments. But the most recent issue of the Annals of Applied Statistics is an exception.

Human lives

Scientists' Elusive Goal: Reproducing Study Results

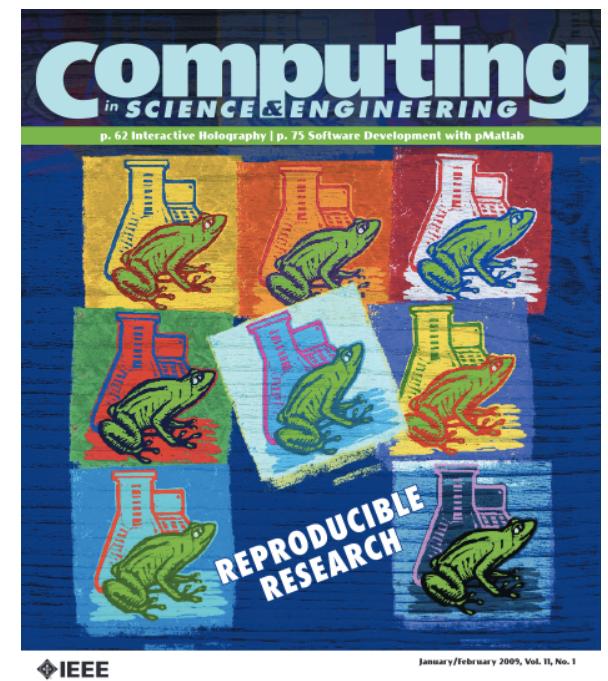


Reproducible Publications and Executable Papers

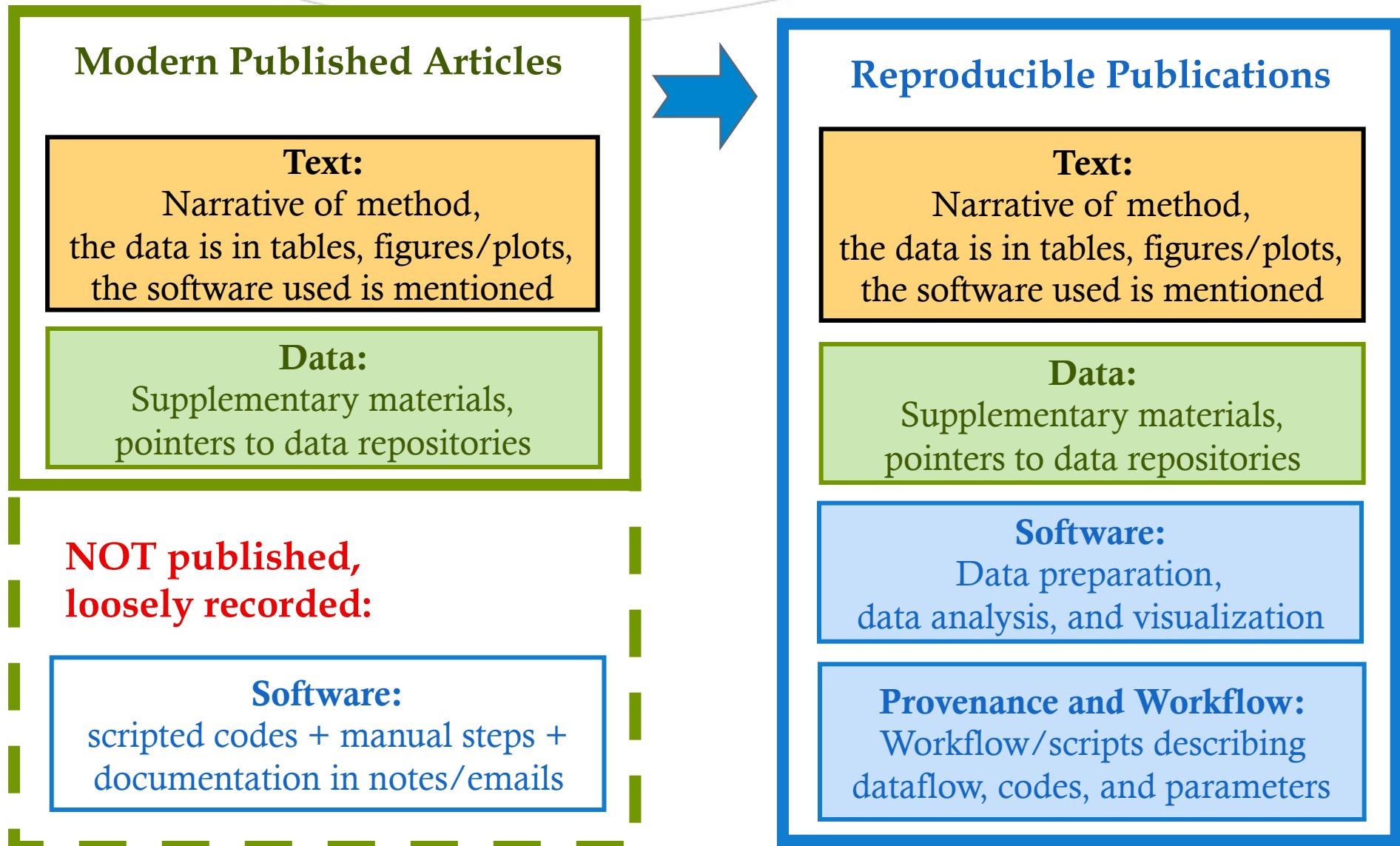


Sweave = R · L^AT_EX

IP[y]: Notebook



Reproducible Articles



Beyond Reproducible Publications

Reproducible Publications

Text:

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

Data:

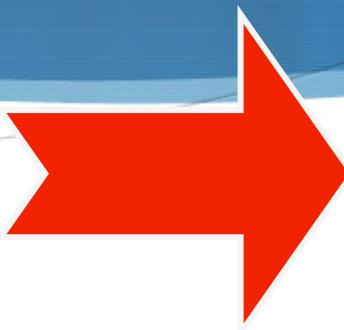
Supplementary materials,
pointers to data repositories

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



Citations: Getting Credit

OPEN  ACCESS Freely available online



Sharing Detailed Research Data Is Associated with Increased Citation Rate

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

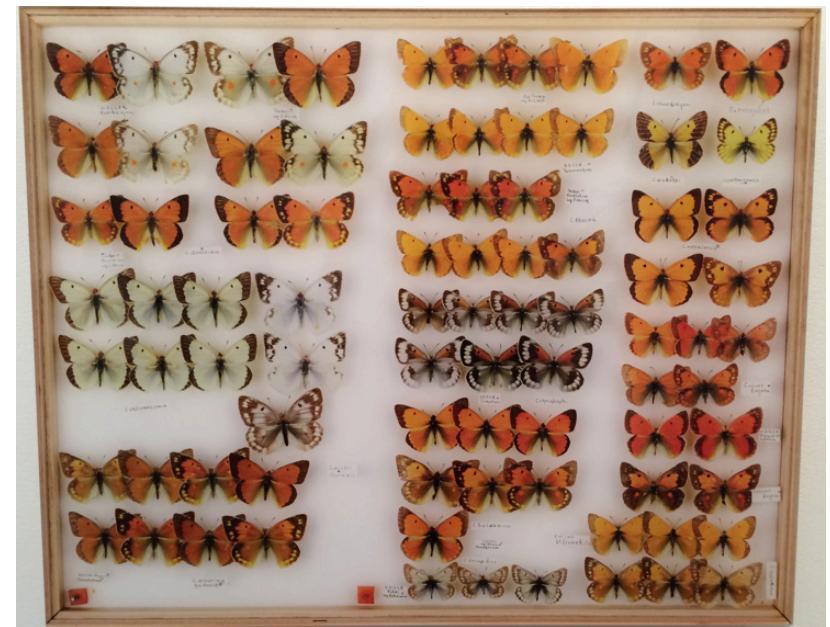
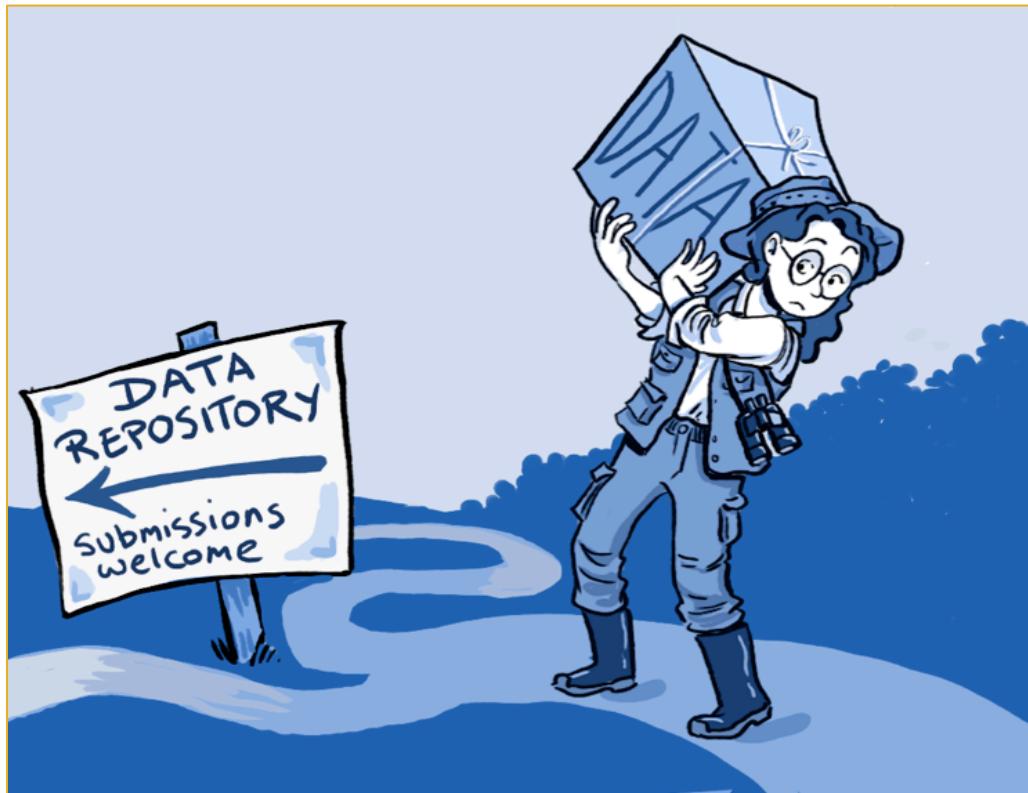
Department of Biomedical Informatics, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, United States of America

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.

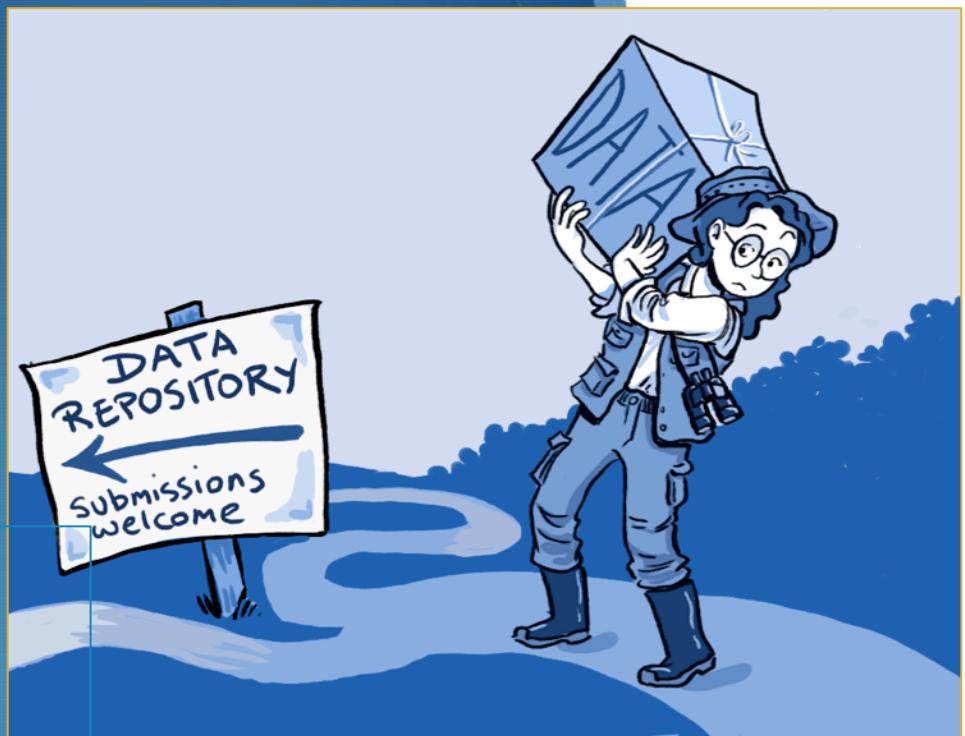
Making Data Accessible

OntoSoft Training

Part 2

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

<http://www.scientificpaperofthefuture.org>



NSF [CER-1440323]
[CER-1343800]

CC-BY
Attribution



"To deposit or not to deposit, that is the question - journal.pbio.1001779.g001" by Roche DG, Lanfear R, Binning SA, Haff TM, Schwanz LE, et al. (2014) - Roche DG, Lanfear R, Binning SA, Haff TM, Schwanz LE, et al. (2014) Troubleshooting Public Data Archiving: Suggestions to Increase Participation. PLoS Biol 12(1): e1001779. doi:10.1371/journal.pbio.1001779.g001.png#mediaviewer/File:To_deposit_or_not_to_deposit,_that_is_the_question_-_journal.pbio.1001779.g001.png

Problems with Current Practice

- ★ Data is often not made available in publications
 - ★ Limited 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^{1*}, Herbert Van de Sompel¹, Robert Sanderson¹, Harihar Shankar¹, Lyudmila Balakireva¹, Ke Zhou², Richard Tobin²

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.



★ Data published in a repository



The US
Long Term Ecological Research
Network

+/-

NTL LTER

"WDNR Yahara Lakes Fisheries: Fish Lengths and Weights 1987-1998" -
Lathrop

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 Å NTL-LTER program Å took over sampling of the Yahara Lakes. However, WDNR data collected from 1997-1998 Å (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 biomonitoring 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_creel_survey_lengths_weights.csv](#)
- [wdnr_creel_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

figshare.com/

Highly connected drug file

Tretinoin	257	46
Levothyroxine	173	36
Methotrexate	156	32
4-Hydroxytamoxifen	115	
Estradiol	98	20
Amantadine	79	1
Rifampin	78	13
Raloxifene	75	18
Propofol	54	5
Indinavir	51	14
Penicillamine	44	10
Daunorubicin	44	12
Tricosan	42	5
Darunavir	40	15

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

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 minimal metadata

3 Domain metadata

4 Unique persistent identifier (PID)

5 Citation preference

Best Practices (1 of 5)

 figshare.com

Highly connected drug file

Published on 20 Aug 2013 - 12:44 (GMT)
Filesize is 4.96 KB

Drug	Count	Count	Annotations
Tretinoin	257	46	Rv1155, aroG, Rv1264, mscL, thyX, gmk, glnA1, Rv1
Levothyroxine	173	36	icl, Rv1264, thyX, glnA1, trpD, leuA, blaI, ethR,
Methotrexate	156	32	Rv0223c, lipJ, Rv1264, ephG, blaI, ethR, sigC, b
4-Hydroxytamoxifen	115	25	cyp130, Rv1264, lppX, gpml, ligA, nirA
Estradiol	98	20	TB31.7, Rv1264, mscL, lppX, coaA, pcaa, Rv3676, fabG1,
Amantadine	79	1	mmaA4, bphD, Rv1264, mscl, thyX, lppX, mmaA2, ptl
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, R
Penicillamine	44	10	mmaA4, Rv1264, thyX, lppX, secA1, serA1, Rv3529c
Daunorubicin	44	12	pepD, Rv1264, thyX, ethR, trx2B,
Triclosan	42	5	pknD, pepD, fabH, Rv1941, devB, ppp, ftsZ, cyp12
Darunavir	40	15	

Enlarge to see the rest of the document

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.

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



1 Publication in a shared repository

2 General minimal metadata

Domain metadata

Unique persistent identifier (PID)



Popular Data Repositories

Not Curated



Curated



"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://www.arqhs.com/articulos/ingeniero-inspector.html>

Directories of Research Data Repositories

- <http://www.re3data.org>
- http://databib.org/index_subjects.php
- 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

Credit: Kerstin Lehnert, LDEO, Columbia U.

Best Practices (2 of 5)

figshare.com/

Highly connected drug file

Published on 20 Aug 2013 - 12:44 (GMT)
Filesize is 4.96 KB

Drug	Count	Count	Description
Tretinoin	257	46	Rv1155, aroG, Rv1264, mscL, thyX, gmk, glnA1, Rv1
Levothyroxine	173	36	icl, Rv1264, thyX, glnA1, trpD, leuA, blaI, ethR,
Methotrexate	156	32	Rv0223c, lipJ, Rv1264, ephG, blaI, ethR, sigC, bj
4-Hydroxytamoxifen	115	25	cyp130, Rv1264, lppX, gpml, ligA, nirA
Estradiol	98	20	TB31.7, Rv1264, mscL, lppX, coaA, pcaA, Rv3676, fabG1,
Amantadine	79	1	mmaA4, bphD, Rv1264, mscl, thyX, lppX, mmaA2, ptl
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, trx2B2,
Triclosan	42	5	pknD, pepD, fabH, Rv1941, devB, ppp, ftsZ, cyp12,
Darunavir	40	15	

Enlarge to see the rest of the document

Cite this: Garioj, 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.

Categories

- Computational Biology

Authors

Daniel Garioj
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 minimal metadata

Domain metadata

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/dc/terms/>)

Minimal Metadata

General

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

Choose a License

Screenshot of the Creative Commons "Choose a License" 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: Other / Multiple formats
License mark: HTML+RDFa

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



CF MetaData

ISO 19115



WaterML2.0

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

Best Practices (3 of 5)

1

Publication in a shared repository

2

General minimal metadata

3

Domain metadata

4

Unique persistent identifier (PID)

5

Citation preference

Domain-Specific Metadata

General

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

Domain Specific

- ★ Collection information
- ★ Pre-processing
- ★ Dataset characteristics

Domain data repositories use metadata standards for that domain and guide you to provide the information needed

Manual Accessibility

SEARCHING AND BROWSING METADATA

- ★ [http://figshare.com/articles/
Highly_connected_drug_file/776887](http://figshare.com/articles/Highly_connected_drug_file/776887)

The screenshot shows a figshare article page. At the top, there's a navigation bar with 'figshare' logo, search bar, 'Browse', 'Upload', 'Sign up', and 'Login'. Below the title 'Highly connected drug file' is a large green download button. The main content area displays a table of highly connected drugs with their counts and associated IDs. A yellow arrow points from the 'Highly_connected_drug_file' link in the previous slide to this download button. At the bottom, there's a 'Share this' section with various social media icons, a 'Cite this' section with a DOI, and a 'Description' section explaining the workflow.

DATA

- ★ [http://files.figshare.com/1175525/
highlConnectedDrugs.txt](http://files.figshare.com/1175525/highlConnectedDrugs.txt)

The screenshot shows a text file with a table of highly connected drugs. The columns are labeled 'Drug' and 'Count'. The data is as follows:

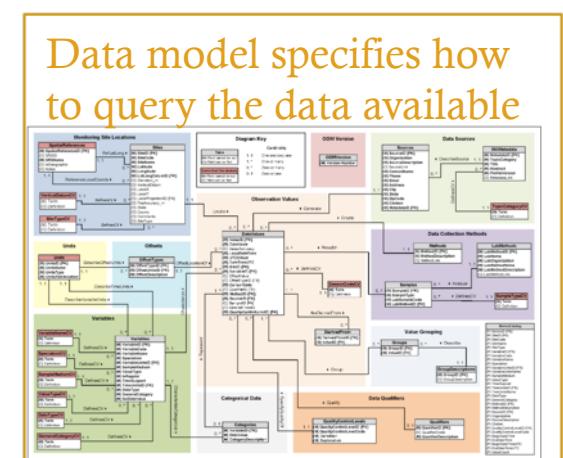
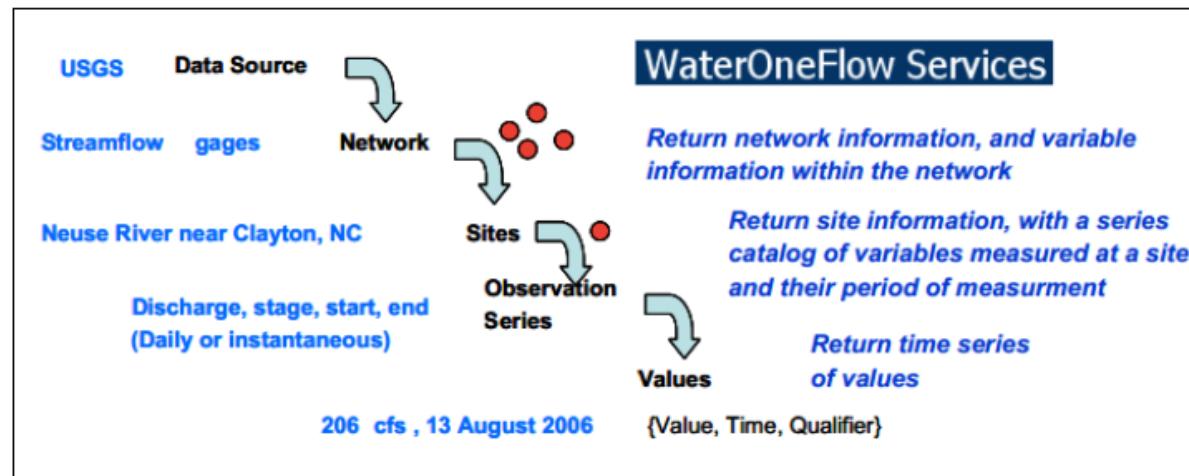
Tretinoin	257	46	Rv1155, aroG, Rv1264, mscL, thyX, gmk, glnA1, Rv
Levothyroxine	173	36	icl, Rv1264, thyX, glnA1, trpD, leuA, blaI, ethR
Methotrexate	156	32	Rv0223c, lipJ, Rv1264, ephG, blaI, ethR, sigC, b
4-Hydroxytamoxifen	115	25	cyp130, Rv1264, lppX, gpm1, ligA, nirA
Estradiol	98	20	TB31.7, Rv1264, mscL, lppX, coaA, pcaA, Rv3676,
Amantadine	79	1	fabG1,
Rifampin	78	13	mmaA4, bphD, Rv1264, mscL, thyX, lppX, mmaA2, pt
Raloxifene	75	18	TB31.7, cyp130, aroG, Rv1264, secA1, trpD, nirA
Propofol	54	5	pth, ethR, clpP, glbN, inhA,
Indinavir	51	14	pknD, lipJ, fabH, Rv1941, Rv3361c, Rv1264, lppX
Penicillamine	44	10	mmaA4, Rv1264, groEL, lppX, secA1, glmU, nusA, R
Daunorubicin	44	12	mmaA4, Rv1264, thyX, lppX, secA1, serA1, Rv3529
Triclosan	42	5	pepD, Rv1264, thyX, ethR, trx2B,
Darunavir	40	15	pknD, pepD, fabH, Rv1941, devB, ppp, ftsZ, cyp12

Machine Accessibility: Metadata is a Necessity!



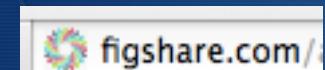
WaterOneFlow Web Services

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.



<https://www.cuahsi.org/Standards>

Best Practices (4 of 5)



Highly connected drug file

Tretinoin	257	46	Rv1155, aroG, Rv1264, mscL, thyX, gmk, glnA1, Rv1264c, Rv1264, thyX, glnA1, trpD, leuA, blaI, ethR, Rv0223c, lipJ, Rv1264, ephG, blaI, ethR, sigC, bj, 25, cyp130, Rv1264, lppX, gpm1, ligA, nirA, TB31.7, Rv1264, mscL, lppX, coaA, pcaA, Rv3676, fabG1, mmaA4, bphD, Rv1264, mscL, thyX, lppX, mmaA2, ptsB, 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, thyX, lppX, secA1, serA1, Rv3529, pepD, Rv1264, thyX, ethR, trx2B, pknD, pepD, fabH, Rv1941, devB, ppp, ftsZ, cyp12:
Levothyroxine	173	36	
Methotrexate	156	32	
4-Hydroxytamoxifen	115		
Estradiol	98	20	
Amantadine	79	1	
Rifampin	78	13	
Raloxifene	75	18	
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 minimal metadata

3

Domain metadata

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
 - Go to <https://w3id.org> (run by W3C), or other PURL services
 - Create a PURL, and direct it to where you actually have the data today eg: [http://www.wisc.edu/
myadvisorsgroup/awesomedata.html](http://www.wisc.edu/myadvisorsgroup/awesomedata.html)
 - Always refer to your data with the same PURL: [http://w3id.org/mydataandme/
awesomedata.html](http://w3id.org/mydataandme/awesomedata.html)
 - Tomorrow you have graduated and tell w3id.org to resolve your PURL to: [http://www.stanford.edu/
myowngroup/awesomedata.html](http://www.stanford.edu/myowngroup/awesomedata.html)
 - It is easy to create your own PURLs, just remember to update whenever you move the data

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 it

Data repositories can issue DOIs for data

DOIs are free

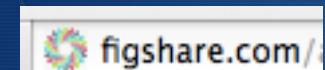
As you may have noticed in the first issue of *PLoS Biology* and again in this issue, there are many places where an alphanumeric string appears after the letters “DOI,” such as [10.1371/journal.pbio.0000005](https://doi.org/10.1371/journal.pbio.0000005) or [10.1371/journal.pbio.0000005.g005](https://doi.org/10.1371/journal.pbio.0000005.g005). 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, Rv1264c, Rv1264, thyX, glnA1, trpD, leuA, blaI, ethR, Rv0223c, lipJ, Rv1264, ephG, blaI, ethR, sigC, bj, 25, cyp130, Rv1264, lppX, gpm1, ligA, nirA, TB31.7, Rv1264, mscL, lppX, coaA, pcaA, Rv3676, fabG1, mmaA4, bphD, Rv1264, mscL, thyX, lppX, mmaA2, ptsB, 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, thyX, lppX, secA1, serA1, Rv3529, pepD, Rv1264, thyX, ethR, trx2B, pknD, pepD, fabH, Rv1941, devB, ppp, ftsZ, cyp12:
Levothyroxine	173	36	
Methotrexate	156	32	
4-Hydroxytamoxifen	115		
Estradiol	98	20	
Amantadine	79	1	
Rifampin	78	13	
Raloxifene	75	18	
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 minimal metadata

3

Domain metadata

4

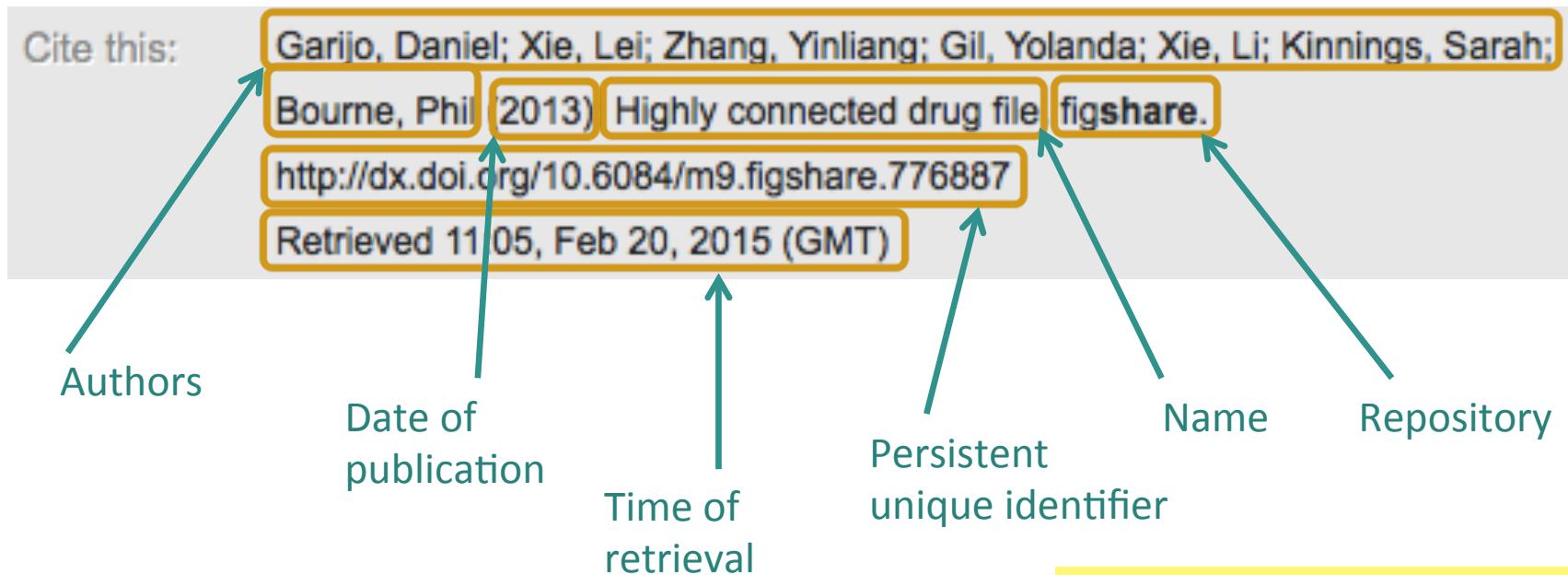
Unique persistent identifier (PID)

5

Citation preference



Data Citation Format



Share this:

[Share](#) 0

[Tweet](#) 0

[g+1](#) 0

[Embed*](#)

Data repositories and journals often specify how to cite data

Versatility of Data Repositories

zenodo

Search 

Upload Communities  gil@isi.edu 

 Delete  Save  Publish

New upload

Instructions: (i) Upload minimum one file or fill-in required fields (marked with a red star). (ii) Press "Save" to save your upload for editing later. (iii) When ready, press "Publish" to finalize and make your upload public.

Upload type required 

 Publication  Poster  Presentation  Dataset  Image  Video/Audio  Software  Lesson

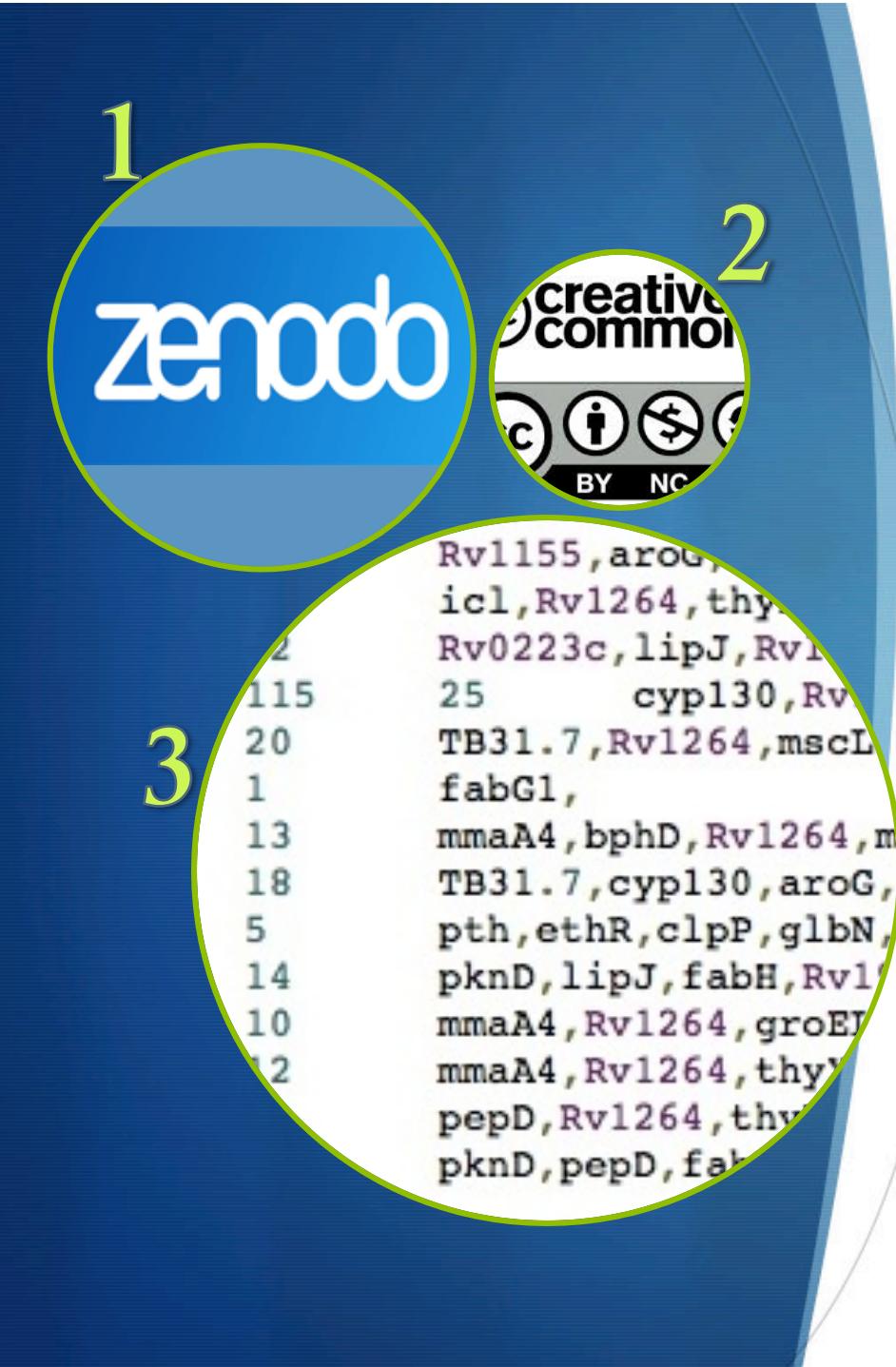
What if...

- ★ ... there are several datasets in several files?
 - ★ Create a DOI for each file and a DOI for the whole set
- ★ ... the data is from a public repository?
 - ★ Publish the query, create a DOI + metadata for it, mention the original source in the metadata, point to the original data source
- ★ ... the data is from a colleague?
 - ★ Get permission in advance and make an agreement, then do as with the data from a public repository
- ★ ... the data comes from many sources?
 - ★ Credit each source, create URIs as needed
 - ★ Can create a table with “microattribution” that summarize each data source
- ★ ... the data comes from a database?
 - ★ Create a file (or files) from it
- ★ ... the data has many versions?
 - ★ Create a DOI either for each slice or for each snapshot



Goals of this Section

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



Making Data Accessible: Simplest Approach

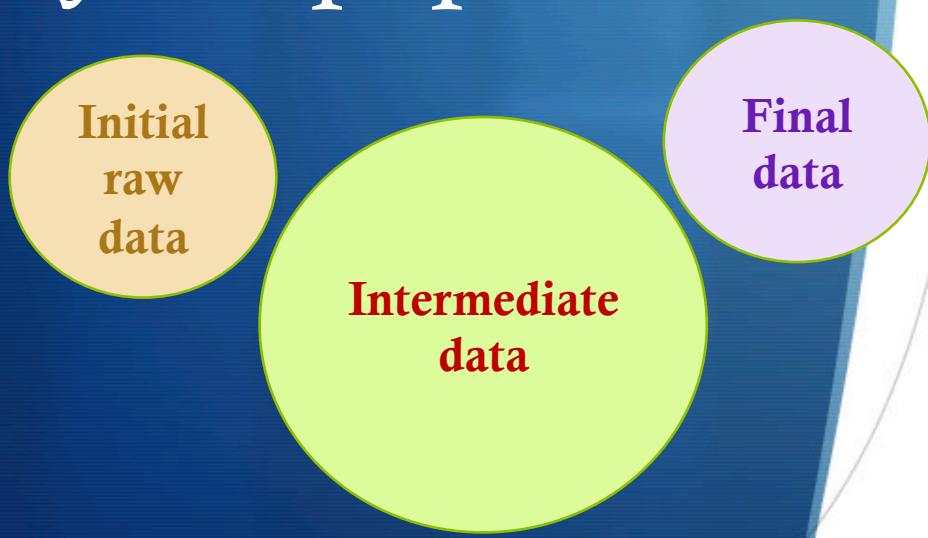
1. Create a public entry for your dataset with a persistent unique identifier
 - Go to a domain repository (use a general repository, e.g., zenodo.org, if you cannot find one), create an account
 - Create an entry for your dataset
 2. Specify the metadata
 - Including license -- choose from <http://www.creativecommons.org/licenses>
 3. Upload/point to the data
- Voilà! The repository will give you a data citation**

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



- ★ 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>

<http://www.flickr.com/photos/gemmerich/6365692623/in/photostream/>



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 other researcher can reproduce the experiments described, (2) our aim to promote openness so that PLOS journals can be built upon by future researchers. Therefore, if new software or a new application 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/>). Using commercial software such as Mathematica and MATLAB does not preclude a paper from being open access, if the software is not preferred.
- **Documentation for running and installing the software.** For end-user applications this may be a simple file; for software libraries, instructions for using the application program interface (API) are a prerequisite; for software libraries, instructions for using the application program interface (API) are a prerequisite.
- **A test dataset with associated control parameter settings.** Where feasible, results should be provided in a standard format. 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 not require users to register for creating user accounts, logging in or otherwise registering personal details. The repository should contain more than 1,000 projects. Examples of such archives are: [SourceForge](#), [Bioinformatics.Org](#), [Savannah](#), [GitHub](#) and the [Codehaus](#). Authors should provide a direct link to the deposited software.

Software Papers and Software Repositories

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



ELSEVIER **software** 
New Journal
Publish your software in SoftwareX
[Find out more](#)



Apache Open Climate Workbench



CSDMS
COMMUNITY SURFACE DYNAMICS MODELING SYSTEM

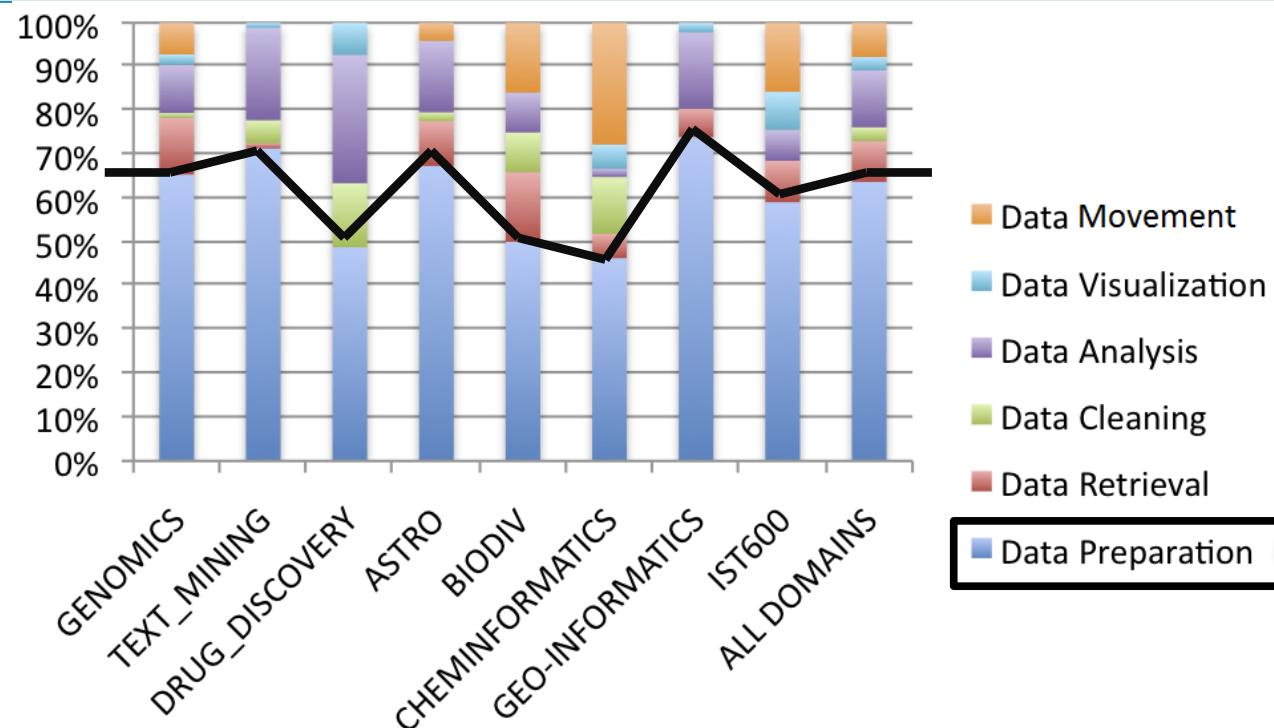
CIG COMPUTATIONAL
INFRASTRUCTURE
for GEODYNAMICS

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, 2013.

“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

Some Notes on Making Software Ready for Publication



- ① Source code vs executable
- ② Making software run elsewhere
- ③ Making software modular
- ④ Making software configurable
- ⑤ Making software report errors
- ⑥ Providing test data
- ⑦ Code analysis

1 Publishing Source Code vs Executable

- ★ **Source code**
 - ★ Improve transparency
 - ★ Opportunity for others to extend the code
 - ★ Software repositories (e.g., GitHub, BitBucket) provide:
 - ★ Version control
 - ★ Community contributions: bugs, extensions, documentation...
- ★ **Executable**
 - ★ Harder to maintain as you have to build for various systems
 - ★ Faster for people to use
- ★ **Important: document run-time dependencies**
 - ★ libraries and other third-party software required both for building the software from source and for running the software

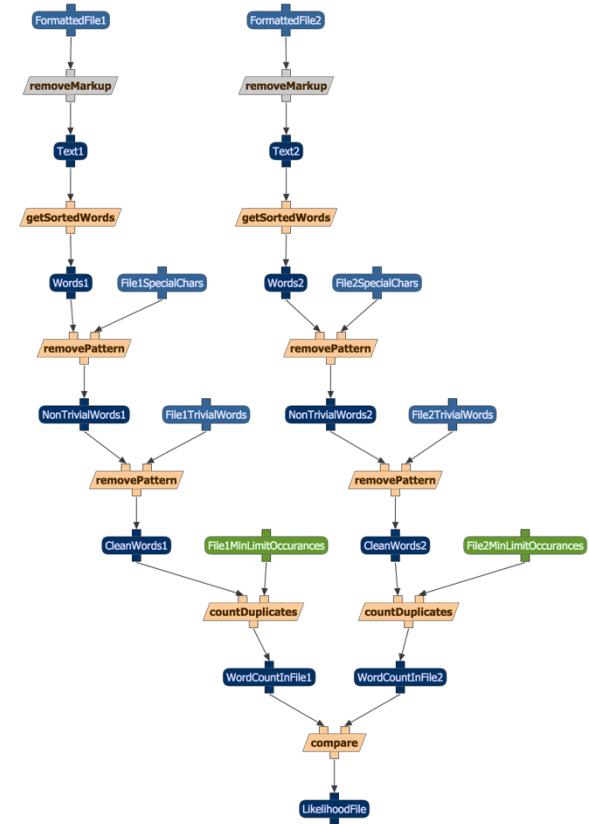
Making Software Run Elsewhere

- ★ **Portability:** If your software runs on machine A, will it run on machine B too?
- ★ List **required dependencies** (software and libraries that are needed to be installed on a machine to get your software to run)
- ★ **No Hard-Coded** machine specific details in the source code
 - ★ Machine specific details such as file location, server name, etc.. should either go in a configuration file, or be provided as a parameter to the software

3

Making Software Modular

- ★ Split code into multiple parts, where each part can be invoked separately if so desired
- ★ Provides finer grained functionality that someone might want to re-use



4

Making Software Configurable

- ★ Expose important parameters in the software in the form of configuration files or parameter values
 - ★ Helps to make your software more useful to people working in slightly different areas but with similar problem scenarios

5 Making Software Report Errors

- ★ Instead of having the code fail silently if something goes wrong:
 - ★ Show an expressive error for users.
 - ★ Return a failure exit code for catching failures from software

Providing Test Data

- ★ Provide test input data, and test results, and instructions on how to run the tests
- ★ Share real input data if possible, and explain data formats
- ★ Provide information about how to create (or where to get) new input data, and what to do with the result data

Code Analysis

- ★ Automated tools are available for many languages
 - ★ Code profilers
 - ★ Code analyzers
- ★ “Code review” sessions among colleagues to critique each other’s code

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)
 - ★ Very involved, but very beneficial if you have many users

Publishing Software in a Code Repository

The screenshot shows a GitHub repository page for 'jihyunoh / GPF'. A large orange callout bubble points from the right side towards the top center of the page, specifically targeting the 'Version Control' section. The 'Version Control' section is highlighted with a red box and contains the following text: '5 commits', '1 branch', '1 release', and '1 contributor'. Below this, there's a commit history table with the following data:

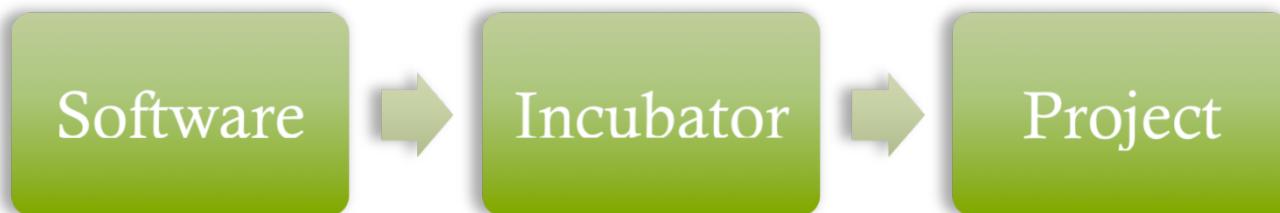
File	Commit Message	Date
LICENSE	Initial commit	3 months ago
README.md	Update README.md	3 months ago
dudt.ncl	add all ncl	3 months ago
dudt_runave.ncl	add all ncl	3 months ago
fv.ncl	add all ncl	3 months ago
grib2netcdf.csh	grib2nc	3 months ago

On the right side of the page, there's a sidebar titled 'Community Contributions' which includes links for 'Code', 'Issues', 'Pull requests', 'Wiki', 'Pulse', 'Graphs', and 'Settings'. At the bottom right, there's a link to the HTTPS clone URL: <https://github.com/jihyun>.

Publishing Software through a Software Foundation



- ★ Oversight of software projects
- ★ Characterized by a collaborative, consensus-based development process and an open and pragmatic software license.



- ★ Apache Software Foundation:
 - ★ <https://www.apache.org/foundation/how-it-works.html#structure>

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

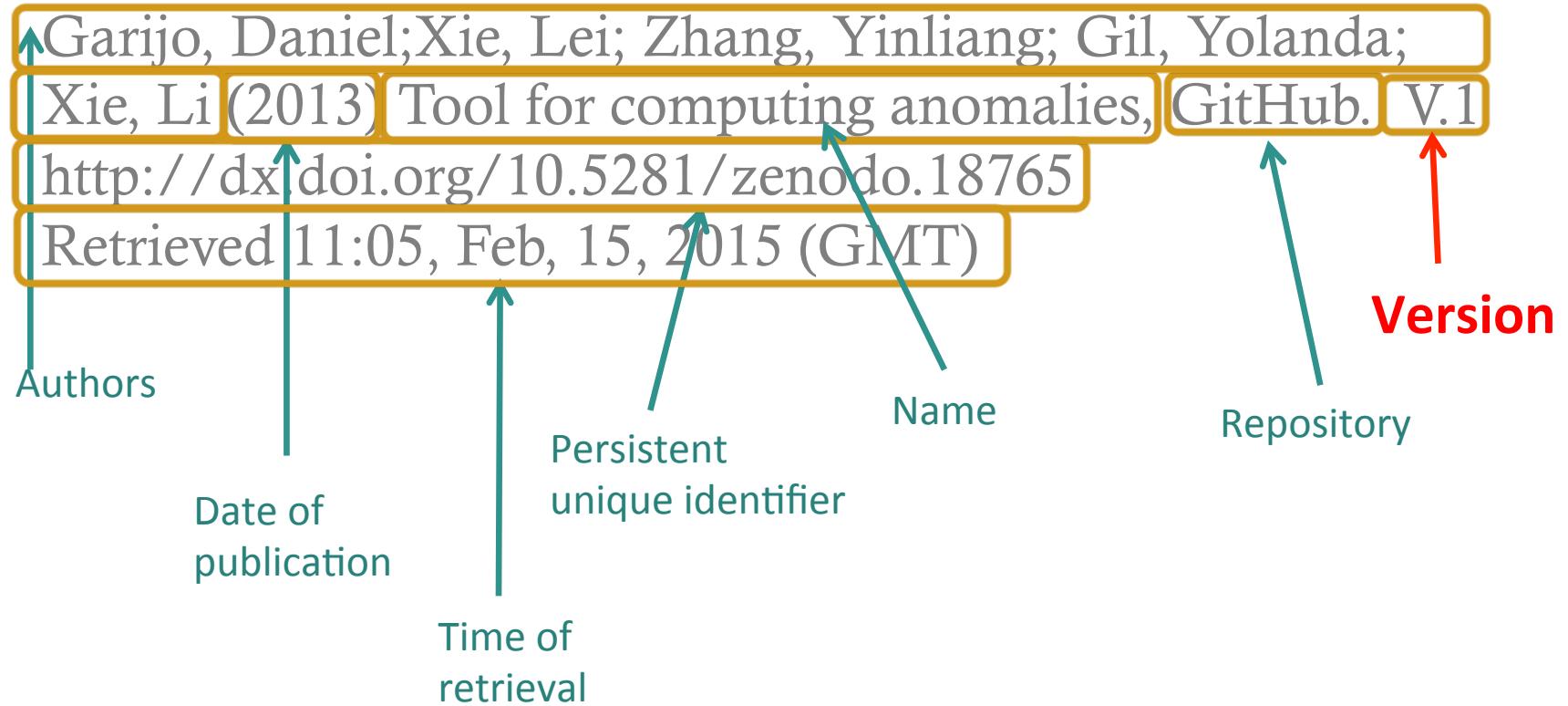


Software Citation

- ★ 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

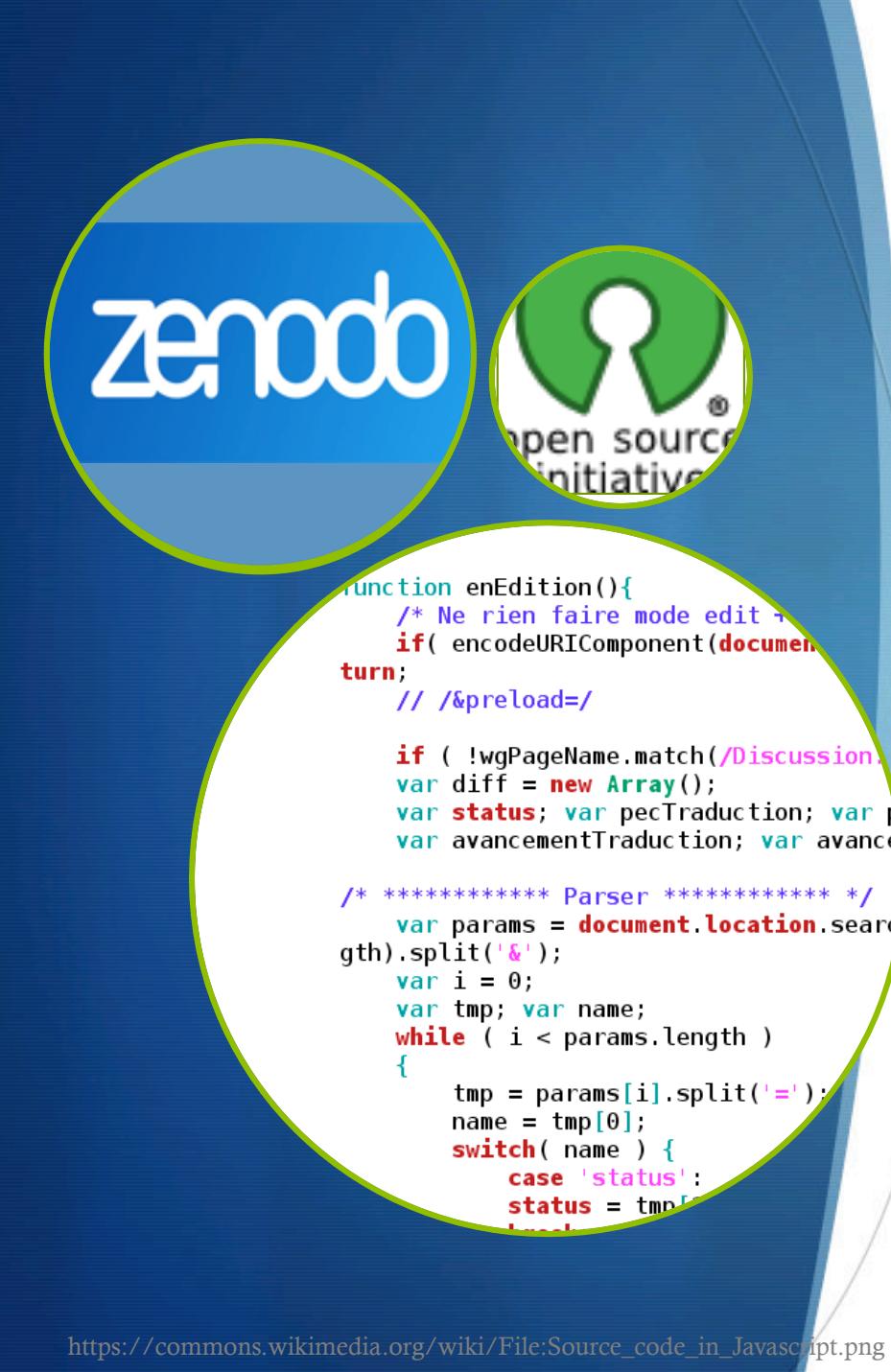
- ★ Similar to data citation format, but includes software version





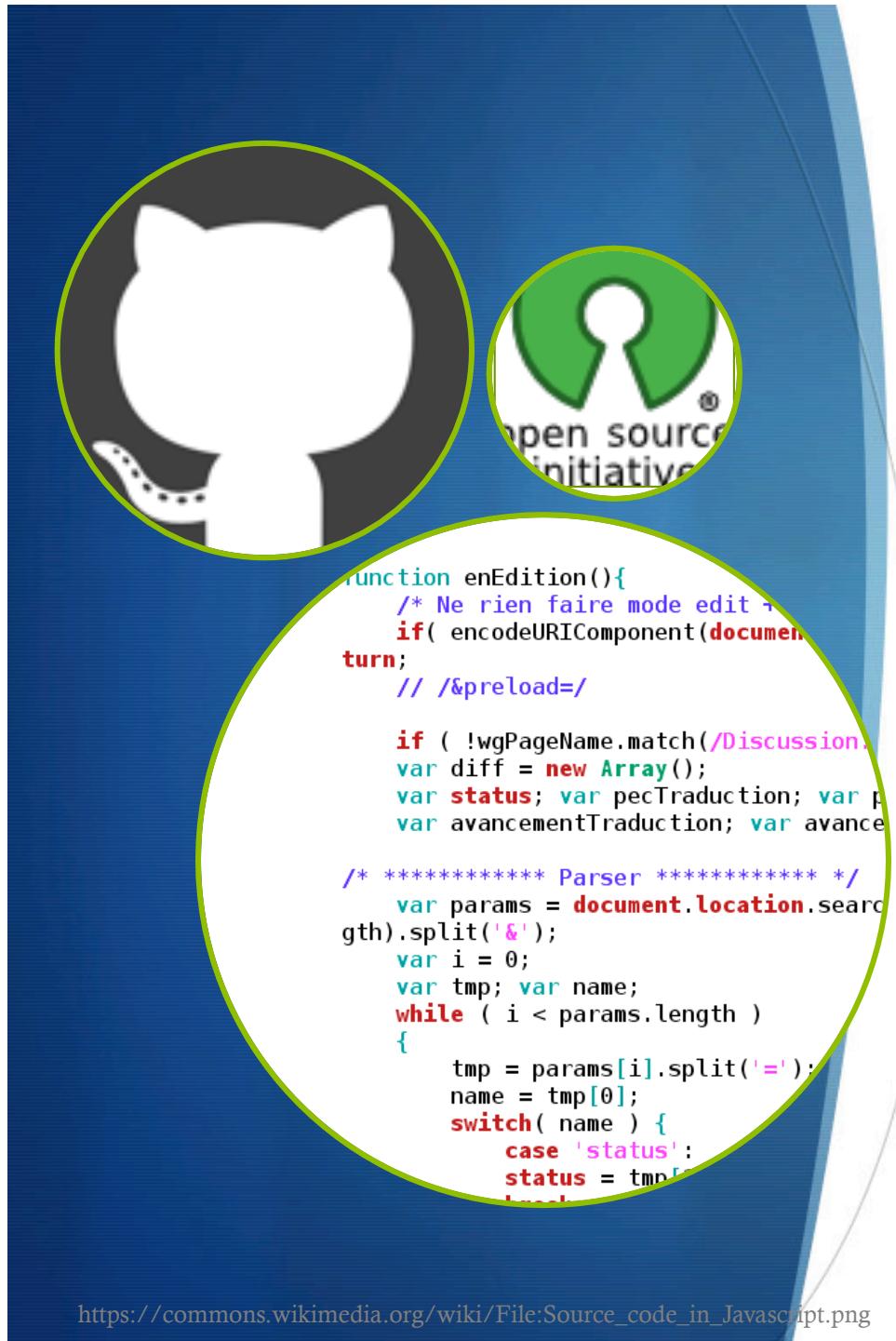
Goals of this Section

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



Making Software Accessible: Ideal Approach

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



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

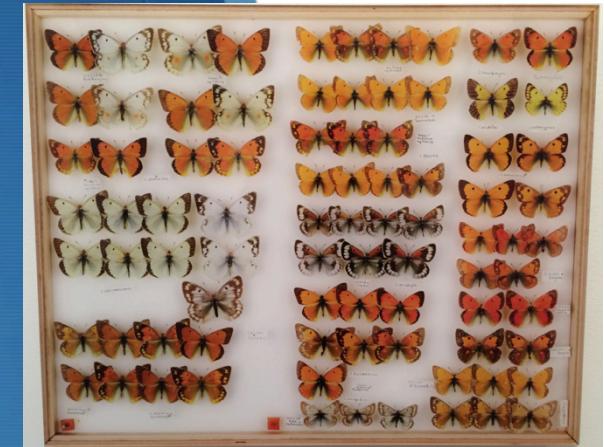
Documenting Software through Metadata

OntoSoft Training

Part 4

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

<http://www.scientificpaperofthefuture.org>



CC-BY
Attribution





Software Repositories

So you have published
your software in a
repository...

Is that sufficient for
others to reuse it?

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





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 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>

OntoSoft Metadata Categories



Software Repository

Describe your software so others can find it

Software List

CSDMS 1D Hillslope MCMC

The model evolves a 1D hillslope according to a linear diffusion rule [e.g. Roering et al. 1999] for boundary conditions idealised as a gaussian pulse. Baselevel fall through time finds the most likely boundary condition parameters when compared...

Author: Martin Hurst

Posted by: admin at 2015-09-08 08:05

CSDMS 2DFLOWWAVE

2D unsteady nonlinear tidal & wind-driven coastal circulation

Author: Rudy Slingerland

Posted by:

C4P 2SA

A software
Posted by:

3DDY

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

C4P AP

A software
..
Posted by:

Software entries
from distributed
repositories are
readily accessible

Semantic
search

Comparison matrix
of software entries

Filter Software List

Search

Author

Keywords: Hydrology

Language: C++

License: Apache License 2.0

Operating System

Publisher

Review

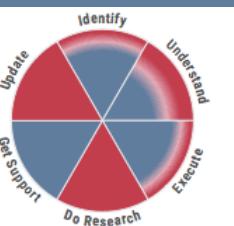
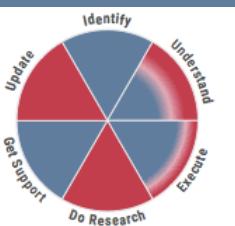
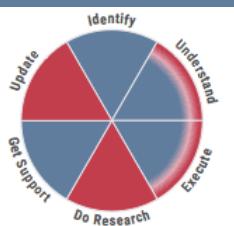
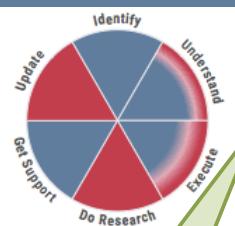
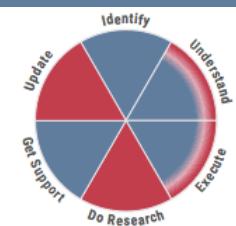
PIHM

PIHMgis

DrEICH

TauDEM

WBMsed



Is there any test data available for the software ?

Metadata
completion
highlighted

Test Data Location:
[http://sourceforge.net/projects/pihmmodel/](http://sourceforge.net/projects/pihmmmodel/)

Test Data Description: Upper Ju
niata River 875 km²: see: [http://sourceforge.net/projects/pihm
model/](http://sourceforge.net/projects/pihm
model/)

Test Data Location:
[http://onlinelibrary.wiley.com
/doi/10.1002
/2013WR015167/full](http://onlinelibrary.wiley.com
/doi/10.1002
/2013WR015167/full)

Test Data Description: Two test
DEMs are included in the reposi
tory, both from Wayne National
Forest.

Test Data Location:
[http://csdms.colorado.edu
/wiki/Model:TauDEM#Testing](http://csdms.colorado.edu
/wiki/Model:TauDEM#Testing)

Test Data Description: The Log
an River DEM is a small test dat
aset useful for learning how to u
se the software

Test Data Location:
[http://csdms.colorado.edu
/wiki/Model:WBMsed#Testing](http://csdms.colorado.edu
/wiki/Model:WBMsed#Testing)

Test Data Description: Extens
ive input dataset is available on th
e CSDMS HPCC (beach) at '/scr
atch/ccny/RGIarchive' and '/sc

What are domain specific keywords for this software ? (eg: hydrology, climate)

Basins, Continental

Basins, GIS

Geomorphology, Hydrological, Be
drock channel erosion

Hydrologically corrected DEM, W
atershed

Sediment flux, Global model, Hy
drological model

What Operating Systems can the software run on ?

Unix
Windows
Linux
Mac OS

Unix
Windows
Linux
Mac OS

Unix
Linux

Unix
Windows
Linux
Mac OS

Unix
Linux

Software is
contrasted
by property

Finding Software

- ★ Any kind of software metadata can be useful to find software
 - ★ “I want R code...”
 - ★ “I want to see software by John Smith...”
 - ★ “I want software that is well supported...”
 - ★ “I want software that simulates water runoff...”
 - ★ “I want software that uses elevation data...”



What if...

- ★ ... there are many versions of the software?
 - ★ Give unique identifiers to the most significant versions that you want to release
 - ★ Relate those versions to one another
- ★ ... the software is already in a public repository?
 - ★ Create a proper documentation and description of the software
- ★ ... the software is relatively small?
 - ★ If you think it may be useful to someone (think of people who do not program!), then release it
- ★ ... the software is a large package with many functions?
 - ★ Consider releasing the large package as a whole for those who want all the functionality
 - ★ Consider also releasing pieces of it with limited functionality that may have a broader audience



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 in a Repository

The screenshot shows a GitHub repository page for the user `jihyunoh / GPF`. The page includes sections for Description, Website, and a detailed commit history. Several metrics and features are highlighted with red circles:

- Description:** A text input field for a short description of the repository.
- Website:** A text input field for a website URL.
- Activity Metrics (Top Bar):**
 - 5 commits
 - 1 branch
 - 1 release (circled)
 - 1 contributor (circled)
- Branch Selection:** A dropdown showing `branch: master`.
- Commit History (Left Column):**
 - `LICENSE` (circled)
 - `README.md` (circled)
 - `dudt.ncl`
 - `dudt_runave.ncl`
 - `fv.ncl`
 - `grib2netcdf.csh`
 - `pgf.ncl`
 - `plot_pgf_x.ncl`
 - `plot_ududx_runave.ncl`
- Metrics (Right Side):**
 - Code:** Includes a link to `Code`.
 - Issues:** 0 issues.
 - Pull requests:** 0 pull requests (circled).
 - Wiki:** A link to the wiki.
 - Pulse:** A link to the pulse dashboard.
 - Graphs:** A link to the graphs feature (circled).
 - Settings:** A link to settings.
- Clone URLs:** Links for cloning the repository via `HTTPS`, `SSH`, or `Subversion`.
- Desktop Clone:** A button to "Clone in Desktop".
- ZIP Download:** A button to "Download ZIP".

Describing Software with OntoSoft

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

The screenshot shows the OntoSoft interface for describing software. At the top, there's a navigation bar with links for OntoSoft, Software, Community, and Training. Below that, a breadcrumb trail says PIHM > Identify > LOCATE. On the left, there's a circular diagram divided into six segments: Identify (top), Understand (top-right), Execute (right), Do Research (bottom-right), Get support (bottom-left), and Update (top-left). A blue button labeled "Locate unique description" is at the bottom of this section. To the right, there are two tabs: "Important" (selected) and "Optional". Under the "Important" tab, there's a question "What is the software called ?" followed by the answer "PIHM". Below it is a question "What is a short description for this software ?" with a detailed answer about PIHM being a multiprocess, multi-scale hydrologic model. Under the "Optional" tab, there's a question "What are general categories (keywords, labels) for this software ?" with three listed: Hydrology, Basins, and Continental. At the very bottom, there's a question "Is there a project website for the software ?" with the answer "http://www.pihm.psu.edu/pihm_home.html".

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

Automatic crawlers import metadata from code repositories (eg GitHub)

Finding Software with OntoSoft

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



Software

Community

Training

Software Repository

Describe your software so others can find and use it

PUBLISH YOUR SOFTWARE

Software List

COMPARE □

▲ Name

DrEICH algorithm

EDIT

PIHM

EDIT

PIHMGis

EDIT

TauDEM

EDIT

WBMsed

EDIT

Filter Software List

Search x

Author

Keywords: Hydrological model
OR Hydrology

Language: C++

License: GNU General Public
License v2.0

GNU General Public Lice x

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

Publishing Software Metadata with OntoSoft

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

The screenshot shows a web-based application for publishing software metadata. At the top left is the OntoSoft logo. On the right is a navigation menu icon. Below the header, the title "PIHM" is displayed, followed by the author's name "[Christopher Duffy]". To the right is a circular rating icon with four segments. Below the title, there are three buttons: "HTML", "RDF/XML", and "JSON", with "HTML" being highlighted and circled in red. To the right of these buttons is a "RATE" button. The main content area is titled "Identify". Under "Identify", there is a section titled "Locate - Unique description". A question "What is the software called ?" has a radio button next to the answer "PIHM". Another question "What is a short description for this software ?" has a list item: "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". A callout box on the right side provides instructions: "Publish metadata as HTML from OntoSoft and add pointer from software repository".

PIHM
[Christopher Duffy]

HTML RDF/XML JSON

RATE

Identify

Locate - Unique description

What is the software called ?

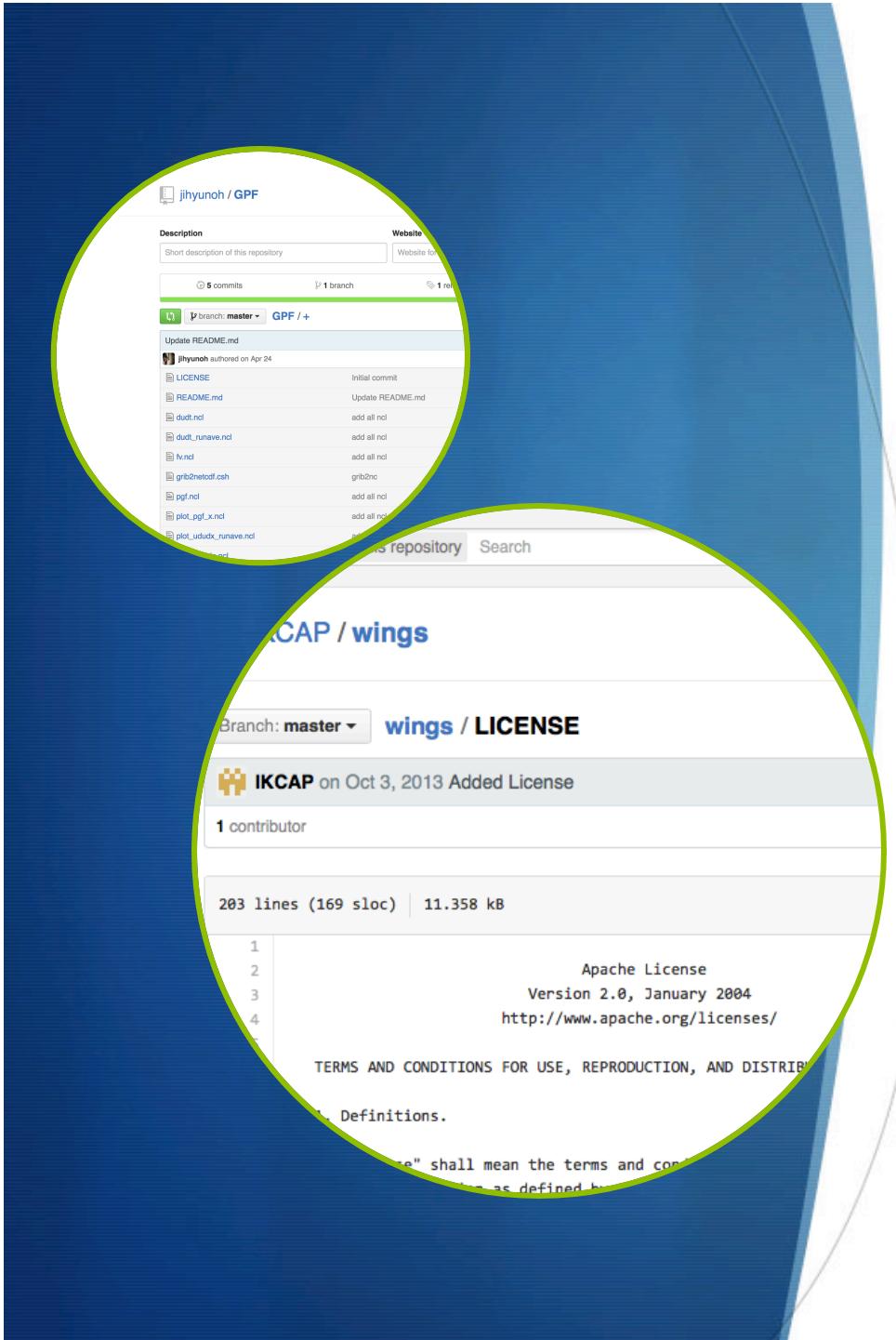
- PIHM

What is a short description for this software ?

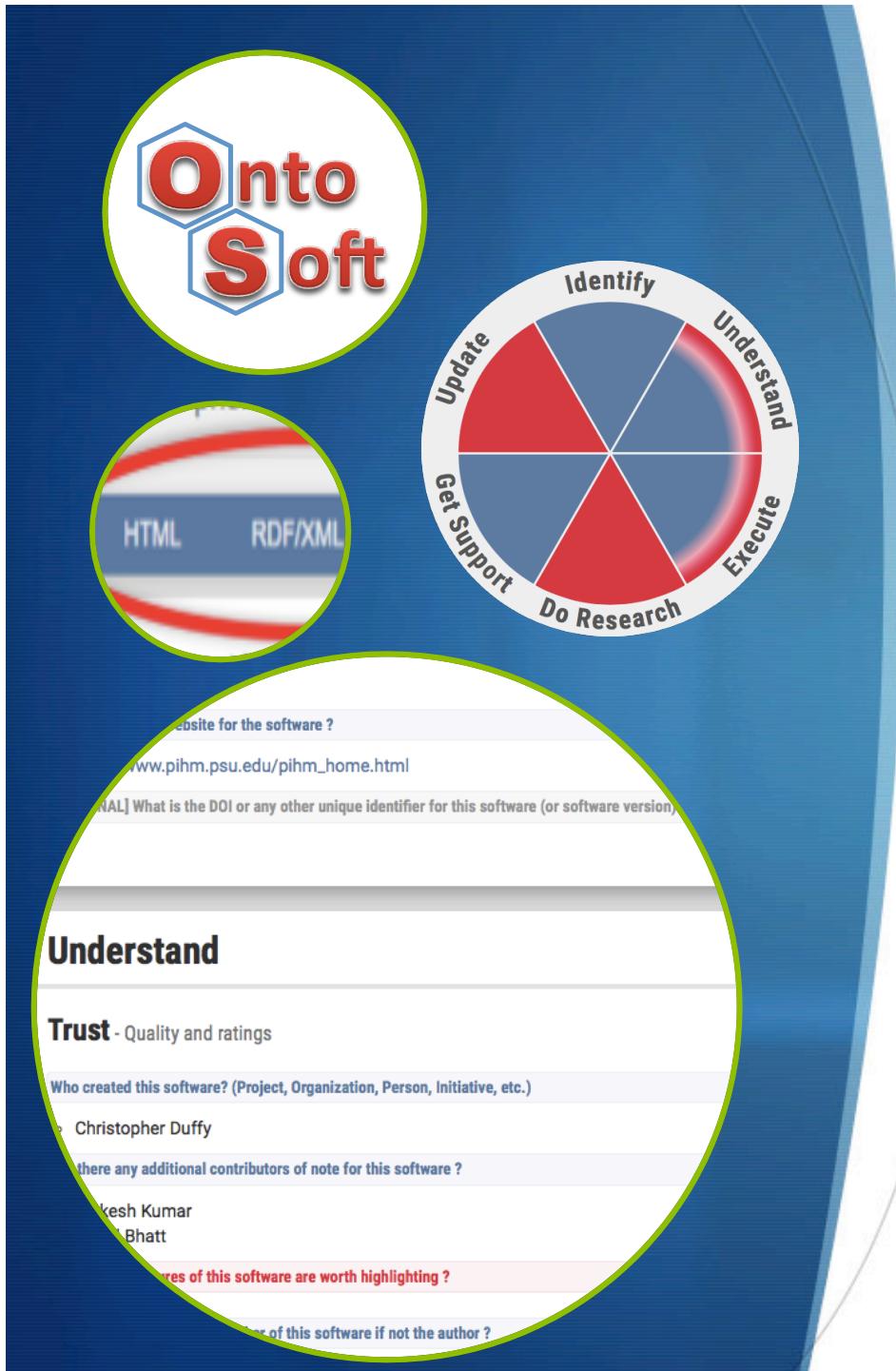
- PIHM is a multiprocess, multi-scale hydrologic model where the major hydrological processes are fully coupled using the semi-discrete finite volume method. PIHM is a physical model for surface and

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



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**

Documenting Provenance and Methods

OntoSoft Training

Part 5

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

<http://www.scientificpaperofthefuture.org>



CC-BY
Attribution



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

[https://www.flickr.com/photos/alterschwede08/3203630740/ \(CC BY-ND 2.0\)](https://www.flickr.com/photos/alterschwede08/3203630740/)

Methods Described in Text Are Incomplete

- ★ 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]

Methods Described in Text Are Ambiguous

- ★ “**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]
- ★ “Ambiguity can occur at the **lexical, syntactic or semantic** level and is not necessarily the result of incompetence or bad practice. It is a natural consequence of using natural language and is unavoidable.” [Ince et al 2012]

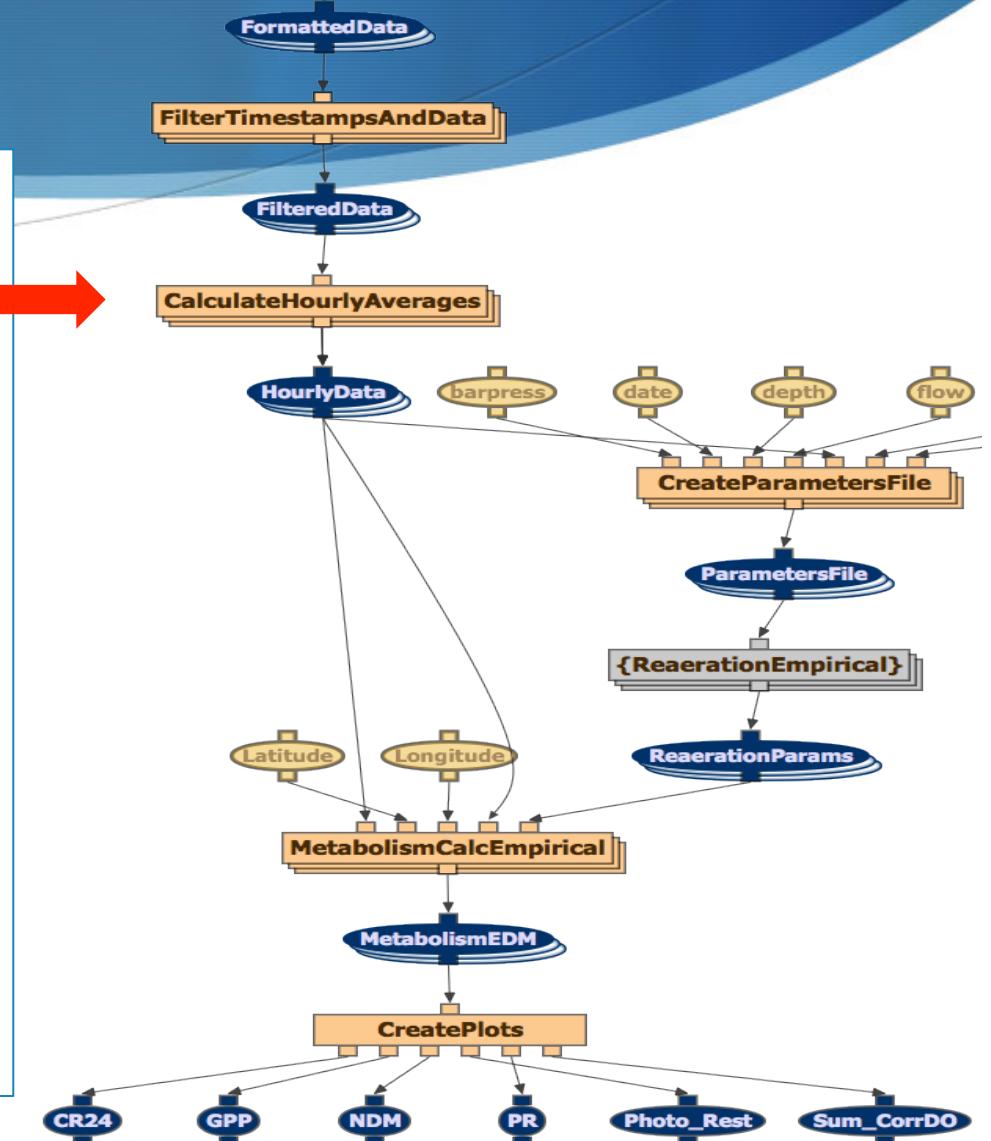


Goals of this Section

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

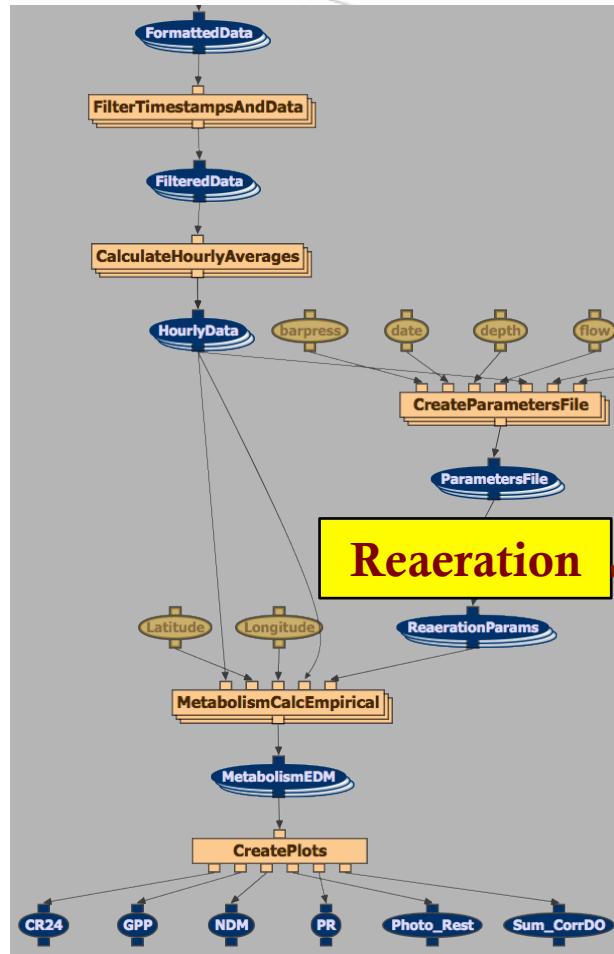
Workflows as Representations of Computational Methods

- ★ 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

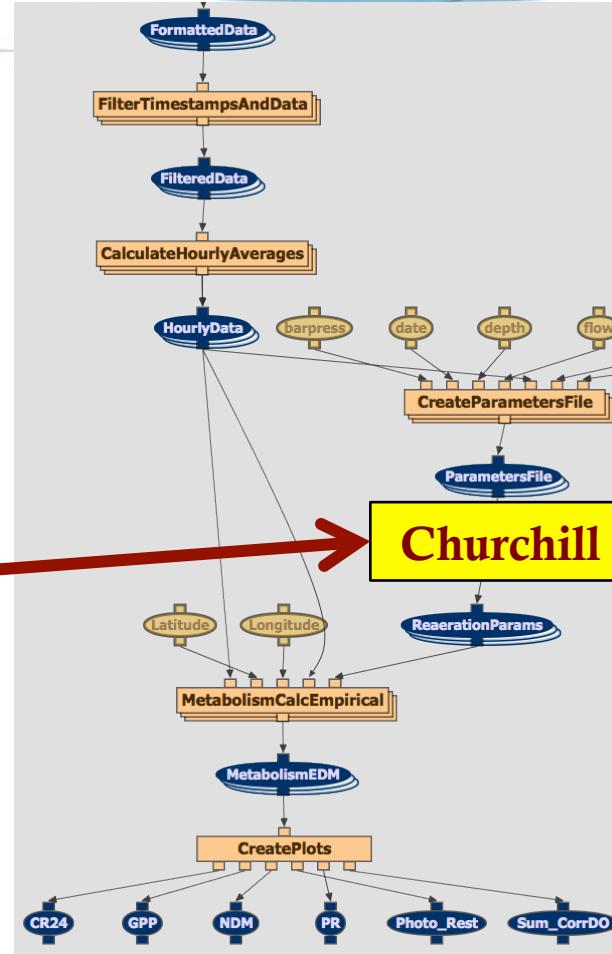


Describing a Method at Different Levels of Abstraction

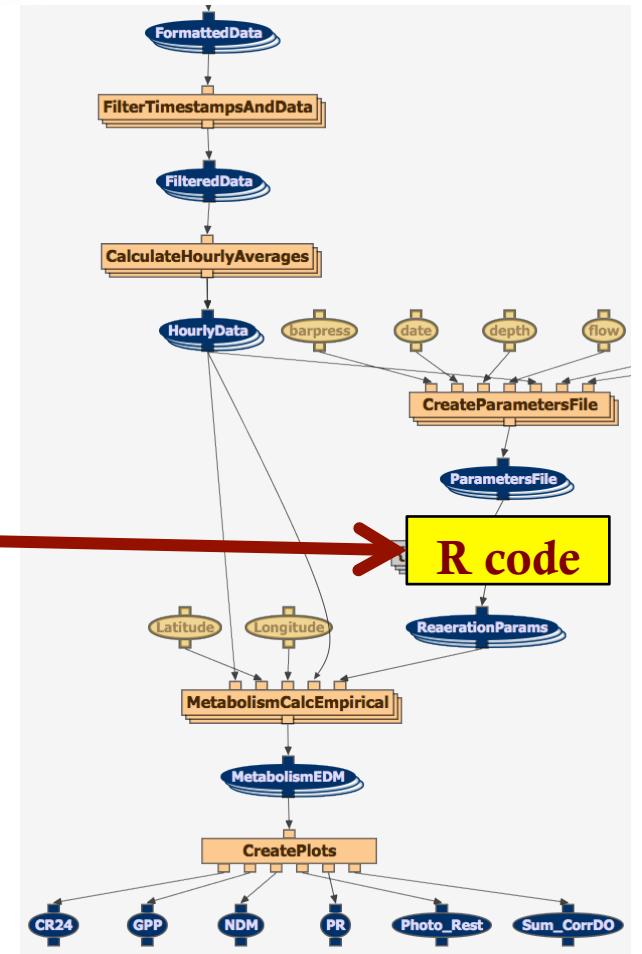
METHODS



ALGORITHMS

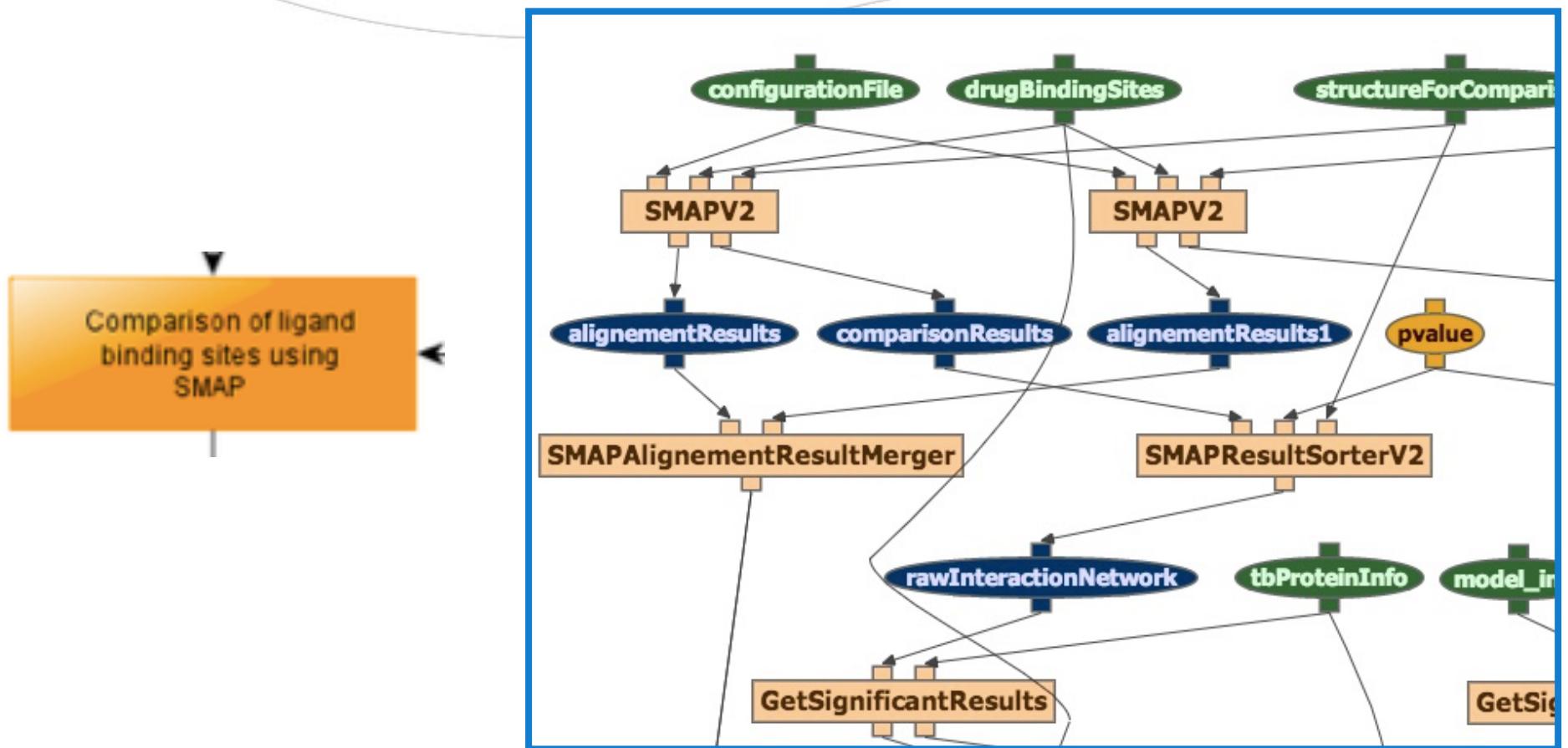


IMPLEMENTATIONS



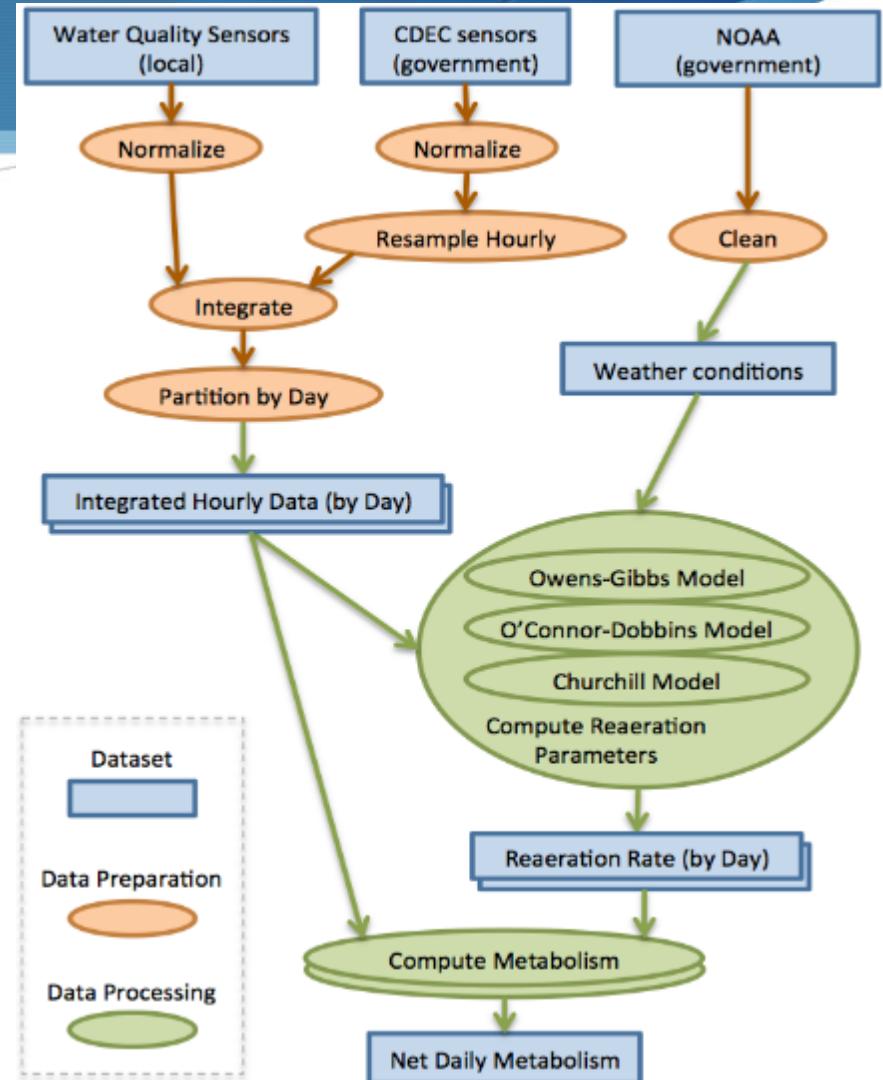
What the Paper Says Versus What the Actual Software Does

(from [Garijo et al 2013])

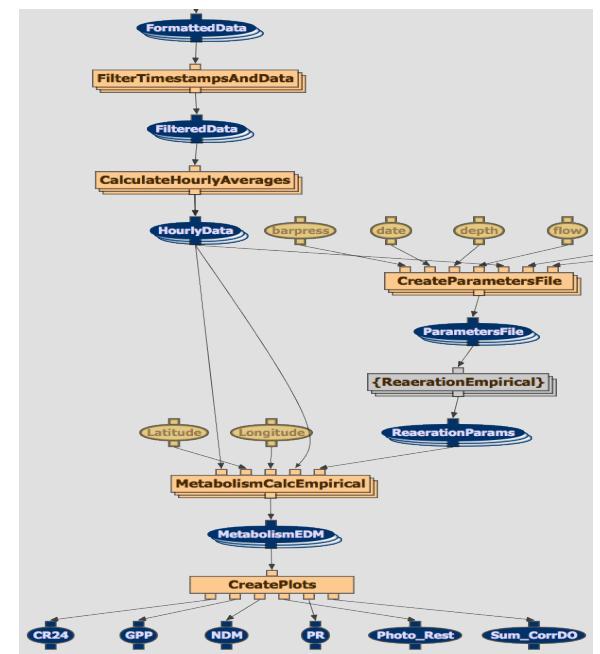
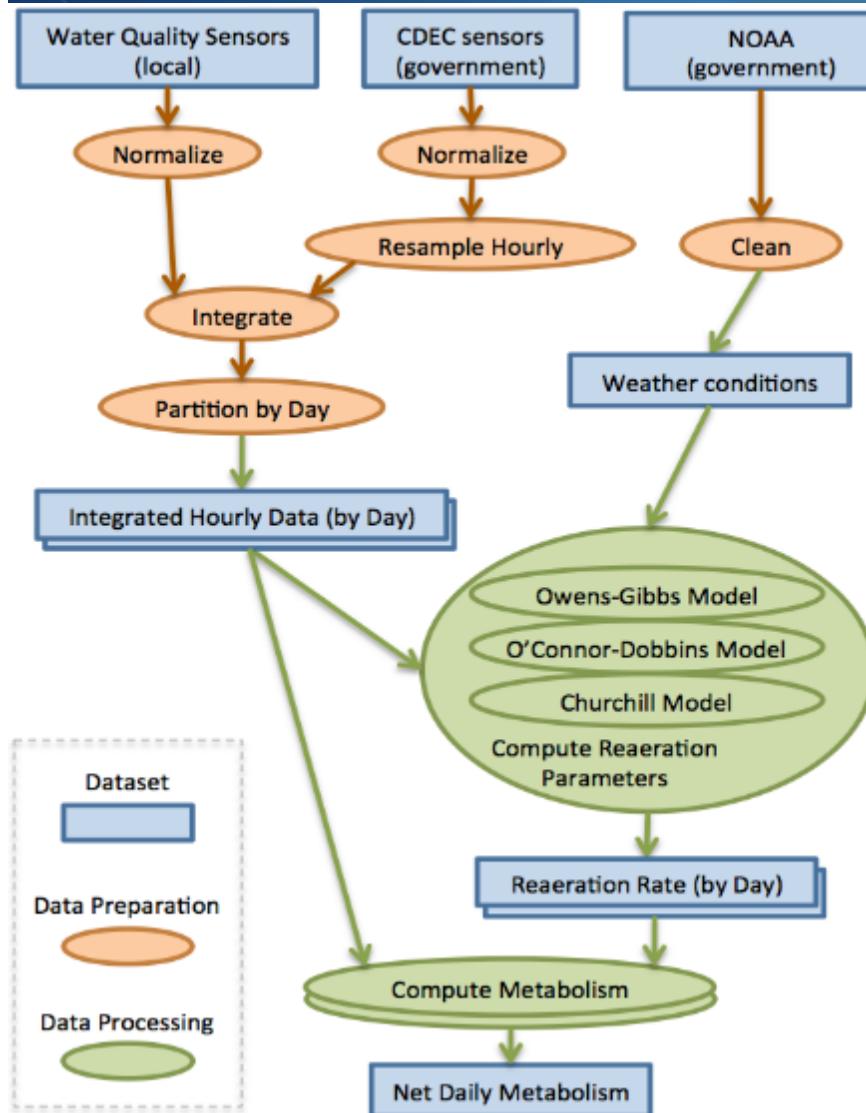


Developing Workflows: How to Sketch a Workflow

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



<http://ipython.org/notebook.html>

IPy IPython Dashboard IPy spectrogram

127.0.0.1:8888/a5222740-848b-4ac1-b212-d732c9f8f78b

IP[y]: Notebook

spectrogram Last saved: Mar 07 11:14 PM

File Edit View Insert Cell Kernel Help

Markdown

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');
```

What is Provenance



Provenance covers:

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

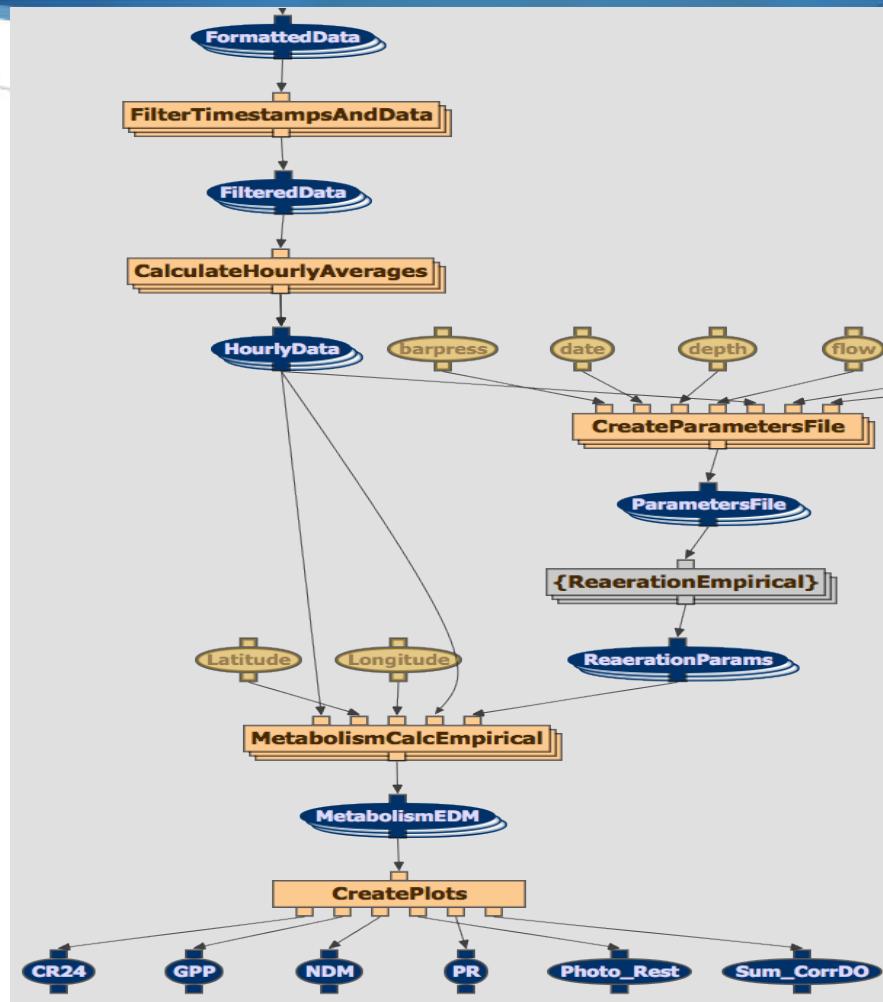
A Working Definition of Provenance

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

1) Provenance as Process (Computing steps, actions, etc)



2) Provenance as Resources (Documents, Data, etc)



WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikimedia Shop

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages

Article Talk

Read

Stratovolcano

From Wikipedia, the free encyclopedia

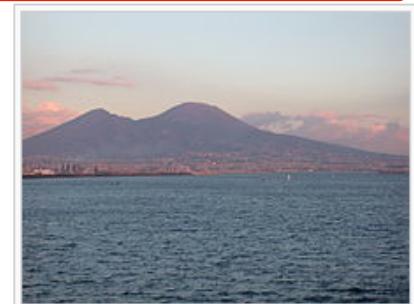
A **stratovolcano**, also known as a **composite volcano**,^[1] is a conical volcano built up by many layers (strata) of hardened lava, tephra, pumice, and volcanic ash. Unlike shield volcanoes, stratovolcanoes are characterized by a steep profile and periodic explosive eruptions and effusive eruptions, although some have collapsed craters called calderas. The lava flowing from stratovolcanoes typically cools and hardens before spreading far due to high viscosity. The magma forming this lava is often **felsic**, having high-to-intermediate levels of silica (as in rhyolite, dacite, or andesite), with lesser amounts of less-viscous **mafic** magma. Extensive felsic lava flows are uncommon, but have travelled as far as 15 km (9.3 mi).^[2]

Stratovolcanoes are sometimes called "composite volcanoes" because of their composite layered structure built up from sequential outpourings of eruptive materials. They are among the most common types of volcanoes, in contrast to the less common shield volcanoes. Two famous stratovolcanoes are **Krakatoa**, best known for its catastrophic eruption in 1883 and **Vesuvius**, famous for its destruction of the towns Pompeii and Herculaneum in 79 AD. Both eruptions claimed thousands of lives.

Existence of stratovolcanoes has not been proved on other terrestrial bodies of solar system^[3] with one exception. Their existence was suggested for some isolated massifs on Mars, e.g., **Zephyria Tholus**.^[4]

References [edit]

1. ^ © This article incorporates public domain material from the United States Geological Survey (USGS) website, specifically the USGS Volcano Hazards Program. Retrieved 2009-01-19.
2. ^ "Garibaldi volcanic belt: Garibaldi Lake volcanic field" (PDF). USGS. 2010-06-27. Retrieved 2010-06-27.
3. ^ Barlow, Nadine (2008). *Mars : an introduction to its interior, surface and atmosphere*. ISBN 9780521852265.
4. ^ Stewart, Emily M.; Head, James W. (1 August 2001). "Volcanoes on Mars". *Geophysical Research Letters* **106** (E8): 17505. doi:10.1029/2000GL000001.
5. ^ a b c d e f g h i j k l m © This article incorporates public domain material from the United States Geological Survey (USGS) website, specifically the USGS Volcano Hazards Program. Retrieved 2009-01-19.



Mount Vesuvius erupted in AD 79 and the last eruption of this stratovolcano near Naples, Italy occurred in March 1944. It has been essentially dormant since then.

3) Provenance as Entities (People, institutions, etc)

Ex: NY Times article from REUTERS reporting “At a press conference last Monday, Buckingham Palace was adamant that Prince Larry did not inhale.”

Title : Prince Larry did not take drugs

Creator : CRETTERS journalist

Subject :

Description :

Publisher : FA Times

Contributor : Duckingham Palace

Date :

Type :

Format :

Identifier :

Source : original CRETTERS article

Language :

Relation : Tapes of the press conference

Coverage :

Rights :

e Larry took drugs

▼ Prince Larry did not take drugs is dismissable

▼ Prince Larry did not take drugs

▼ [more]

according to source **Duckingham Palace** which is completely reliable (A)

and improbable because **They want to save the reputation of the Monarchy**

▼ Prince Larry took drugs is elaborated in [Prince Larry took cannabis](#) and [The trouble with Prince Larry](#)

▼ Prince Larry took cannabis

▼ [more]

according to source **BBC News** which is completely reliable (A)
and confirmed by other sources

► [The trouble with Prince Larry](#)

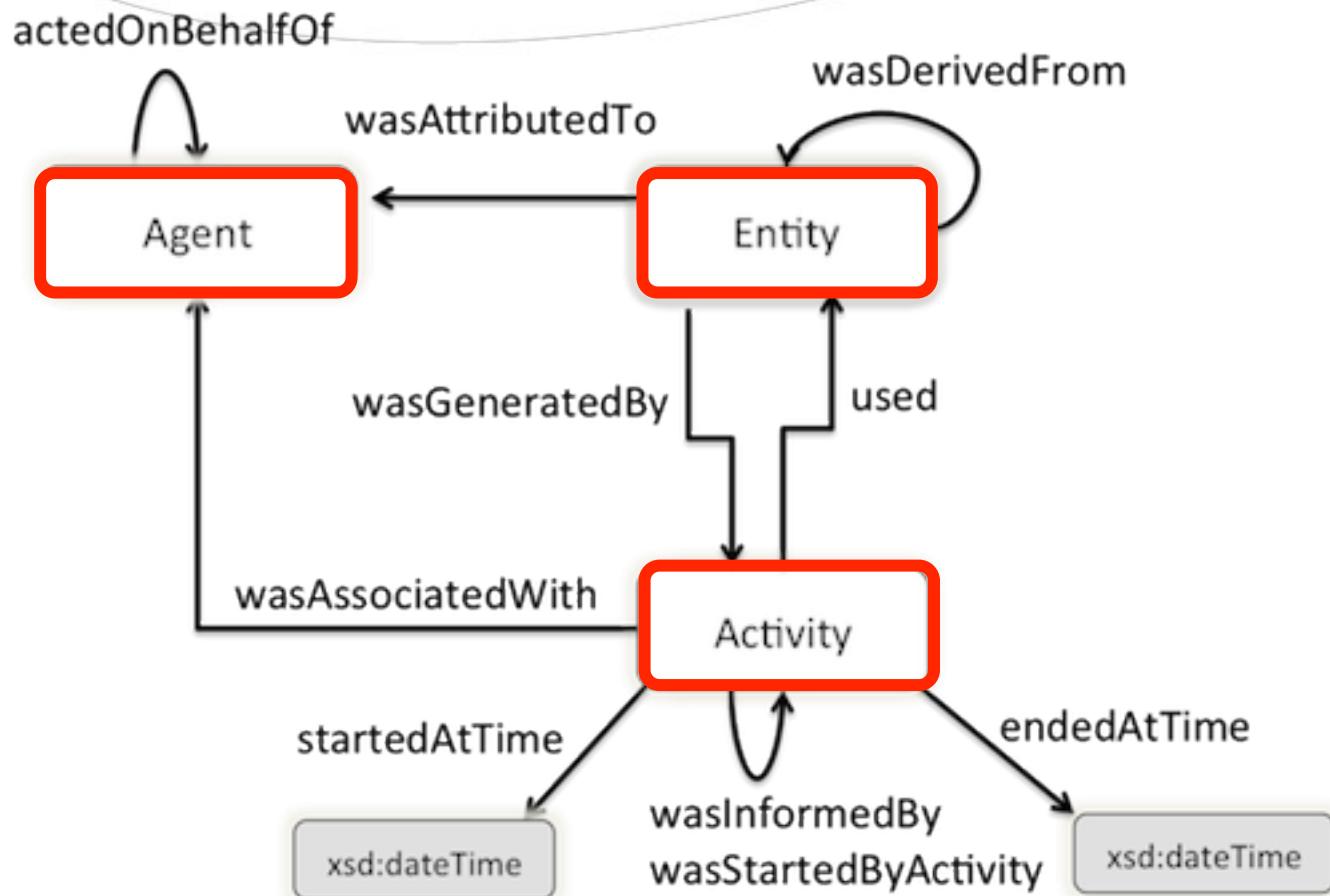
► [more drug problems](#)

A Well-Known Provenance Vocabulary: The Dublin Core

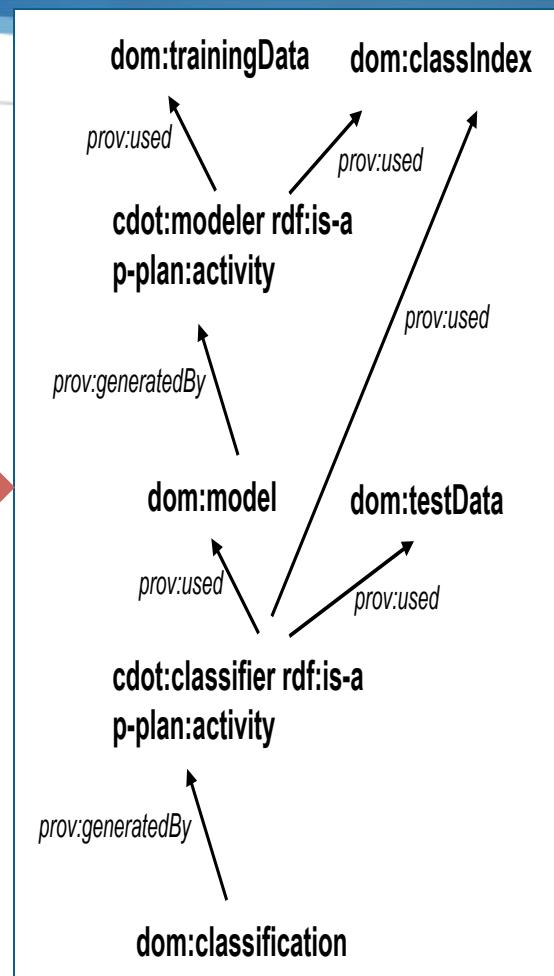
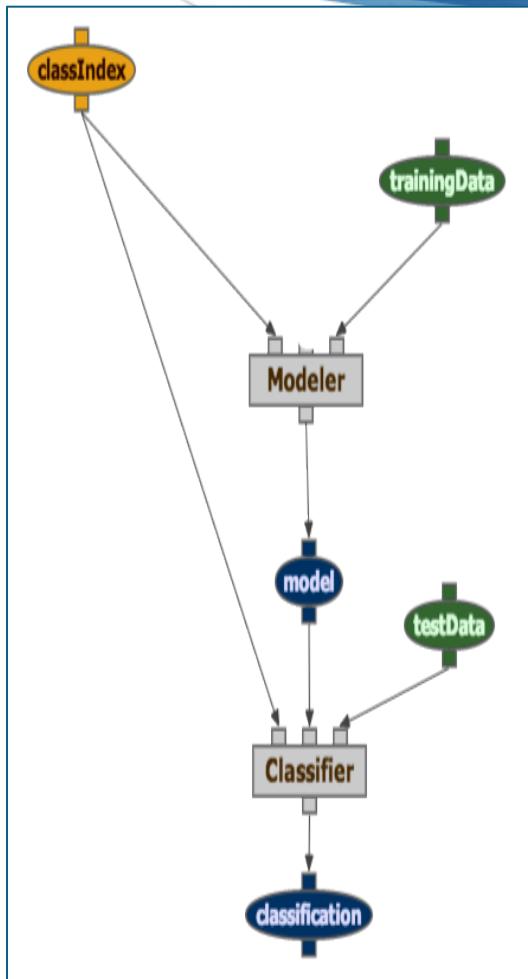
From library sciences

<http://dublincore.org/documents/dcmi-terms/>

A Provenance Standard for the Web: W3C PROV



Representing Provenance with the W3C PROV Standard



Entities

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

Activities

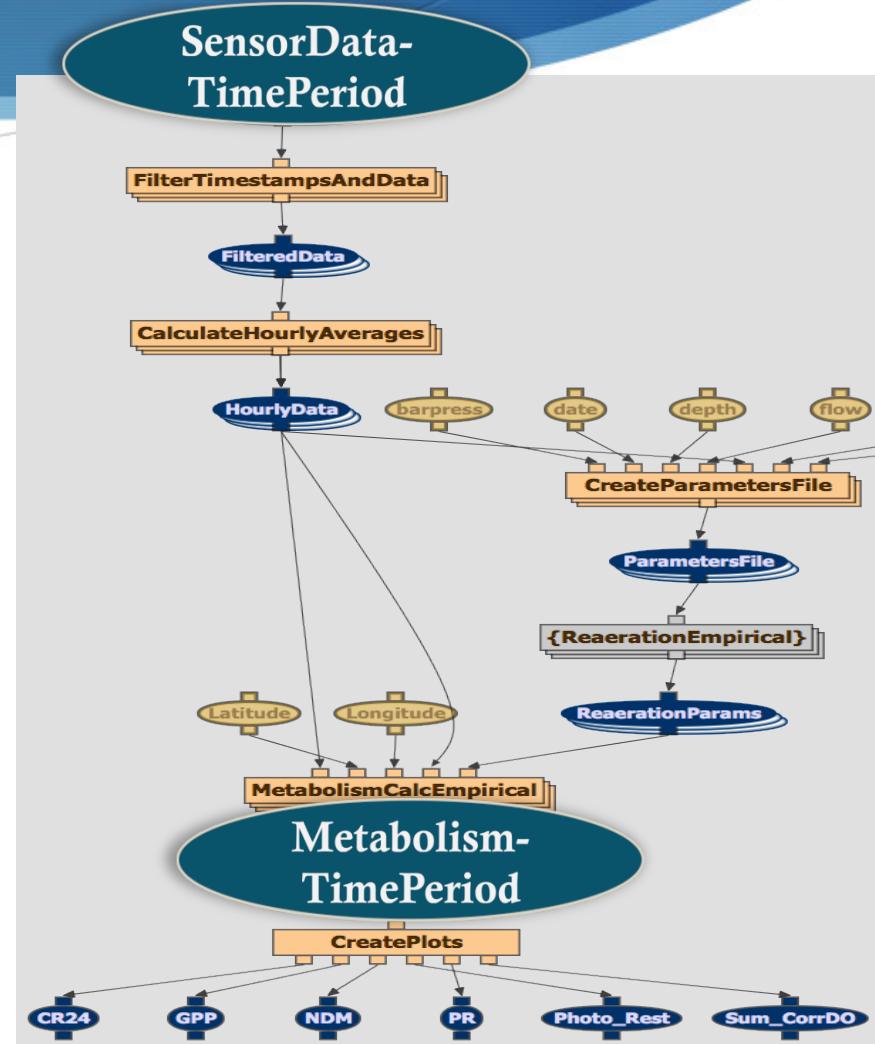
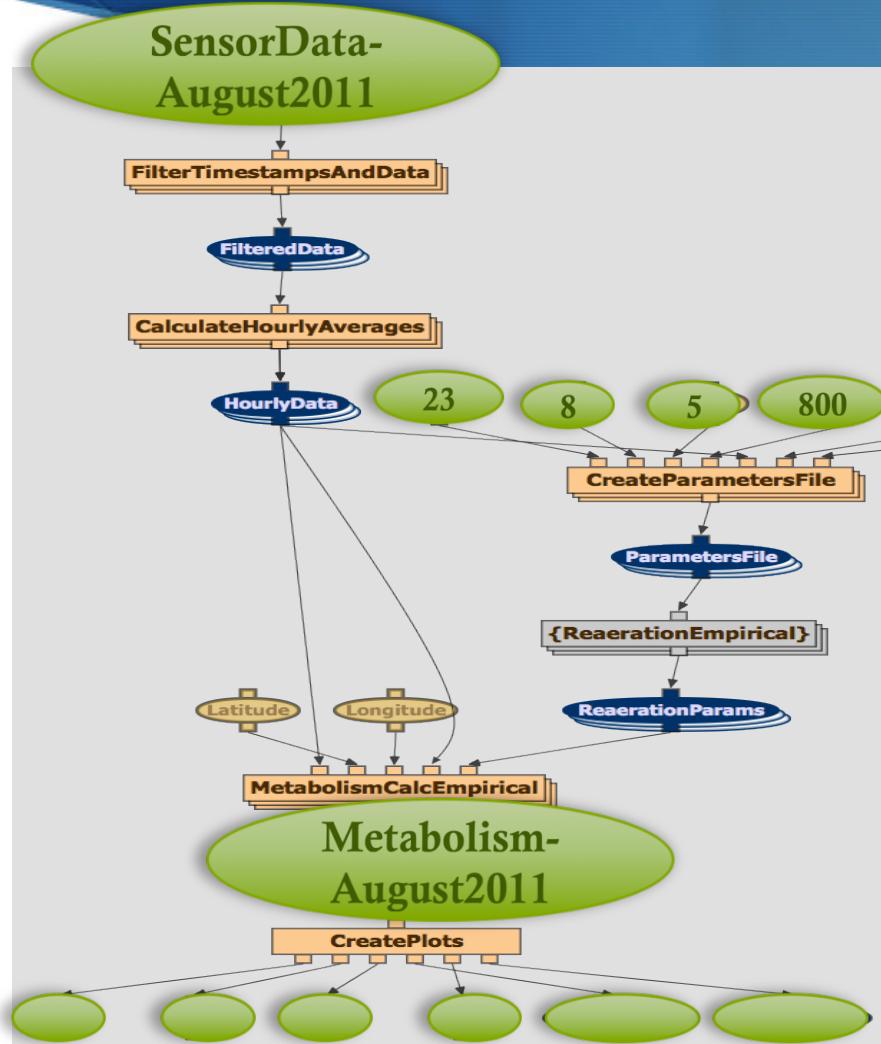
ex:Classifier1 a prov:Activity .

Usage and Generation relations between entities and activities

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

ex:classification1
prov:wasGeneratedBy
ex:Classifier1 .

Describing Execution (Provenance) vs General Method (Workflow)



Publishing Provenance and Workflows

- ★ Hard to deposit workflows or provenance in a repository
 - ★ Not many repositories available
 - ★ Not many communities sharing repositories
 - ★ This will change in the near future
- ★ Publish workflow and/or provenance in a data repository, get a persistent identifier, and cite

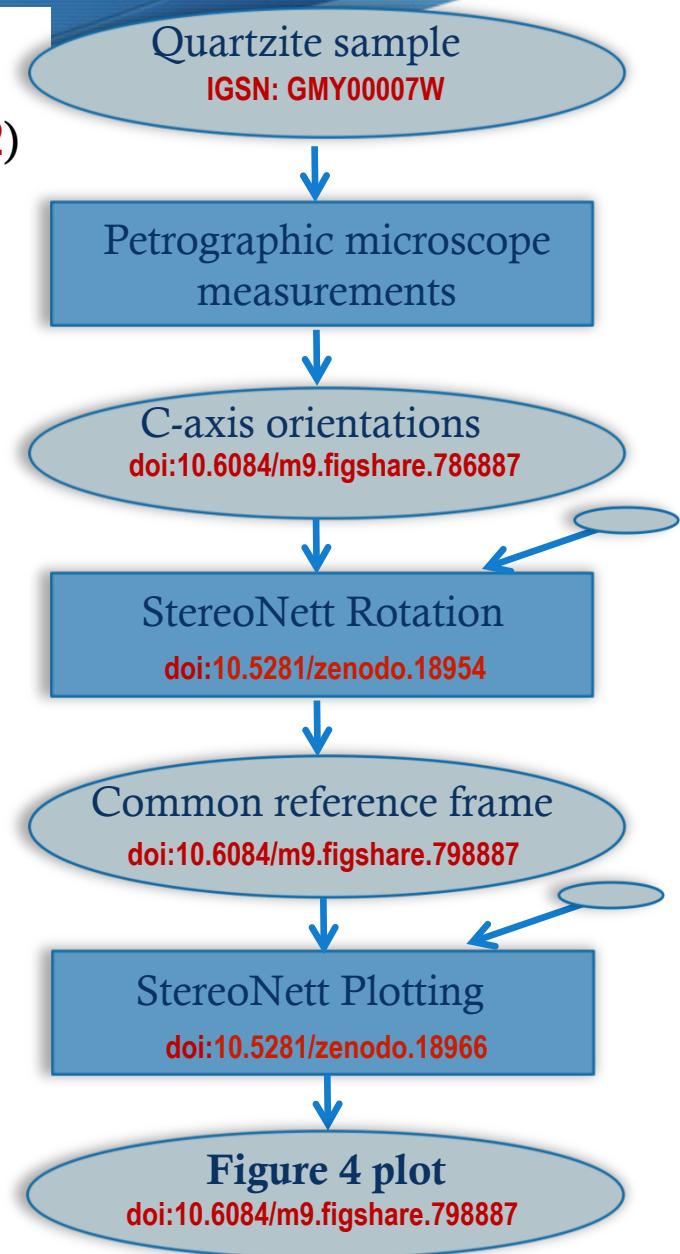


Example: Text and Provenance

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

Jade Silverstein (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)), shown in Figure 4. **The provenance is shown in Fig 5.** [...]





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

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 trace

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 drug was scanned in order to generate a representation of the

2



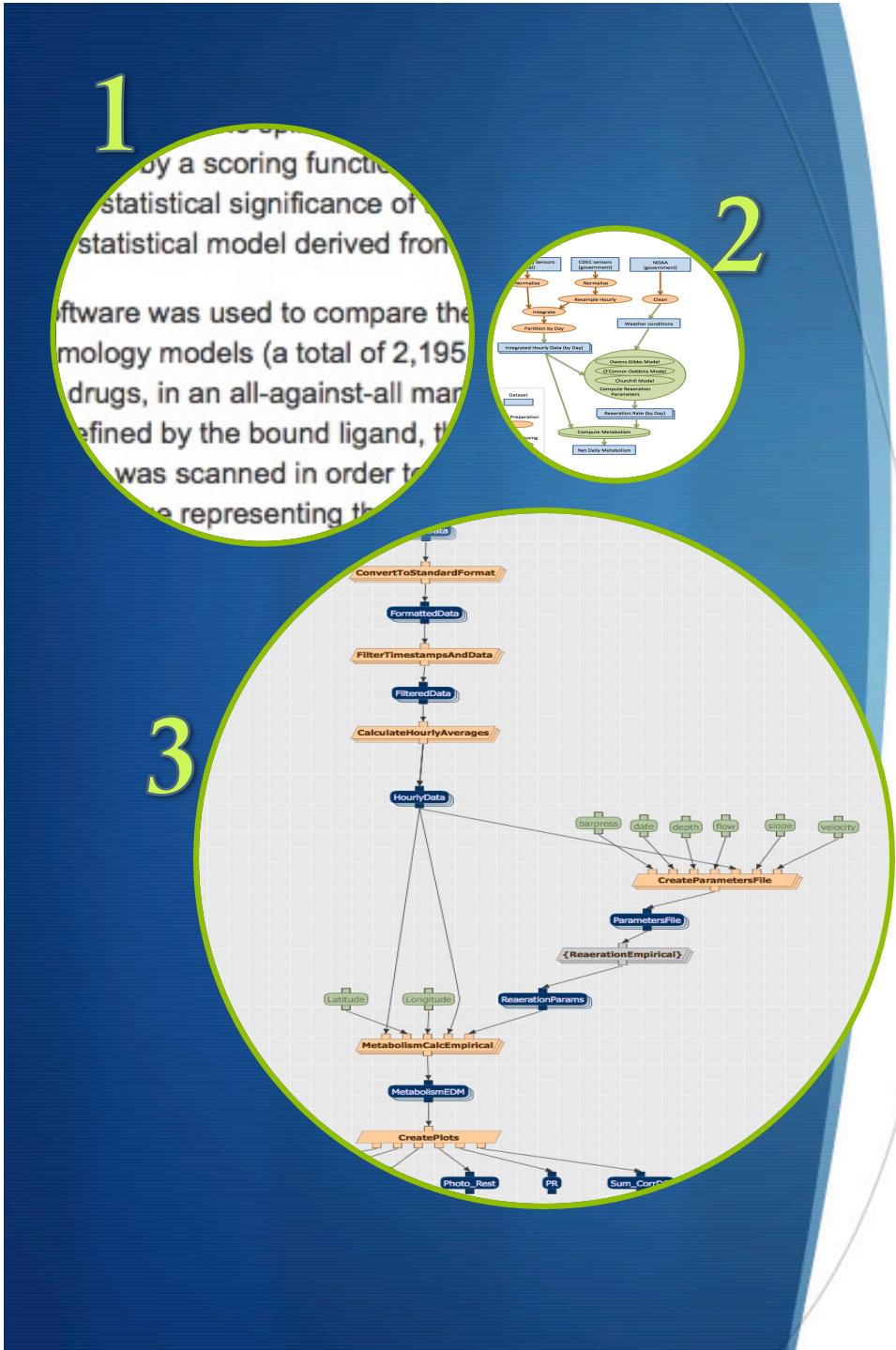
3

```
CreateParametersFileNode_9  
-----  
/usr/share/tomcat6/storage/users/admin/Water/code/library/CreateParametersFile/run -o  
/usr/share/tomcat6/storage/users/admin/Water/data/Params_SMN_2010-03-03Z  
CreateParametersFileNode  
-----  
/usr/share/tomcat6/storage/users/admin/Water/code/library/CreateParametersFile/run -o  
/usr/share/tomcat6/storage/users/admin/Water/data/AvgHourly_SMN_2010-03-03Z  
CreateParametersFileNode_5  
-----  
/usr/share/tomcat6/storage/users/admin/Water/code/library/CreateParametersFile/run -o  
/usr/share/tomcat6/storage/users/admin/Water/data/AvgHourly_SMN_2010-03-03Z  
CalculateHourlyAveragesNode_6  
-----  
/usr/share/tomcat6/storage/users/admin/Water/code/library/CalculateHourlyAverages/run -o  
/usr/share/tomcat6/storage/users/admin/Water/data/
```

Documenting Provenance and Methods:

Ideal 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. 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: Improving Author Citation Profile and Researcher Impact

Part 6

Author Carpentry Training

Gail Clement

<http://www.scientificpaperofthefuture.org>



CC-BY
Attribution



AuthorCarp^{entry}

RETOOLING AUTHORSHIP,
PUBLISHING, IMPACT & CREDIT

Establishing Your ORCID Profile & Researcher Impact

Gail Clement

Head of Research Services, Caltech Library

<http://orcid.org/0000-0001-5494-4806>

gperetsm@caltech.edu



AAAI -17 Tutorial | 4 Feb 2017 | San Francisco

Learn to Write a Scientific Paper of the Future:

Reproducible Research, Open Science, and Digital Scholarship

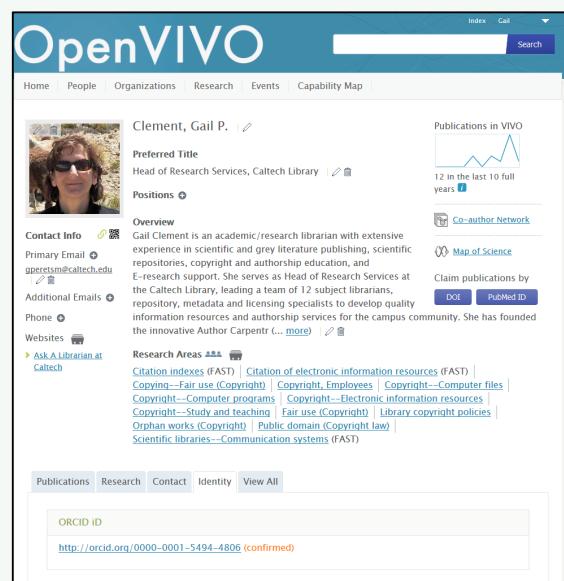


How to build your profile,
establish your reputation,
and get lots of credit
in 5 easy steps

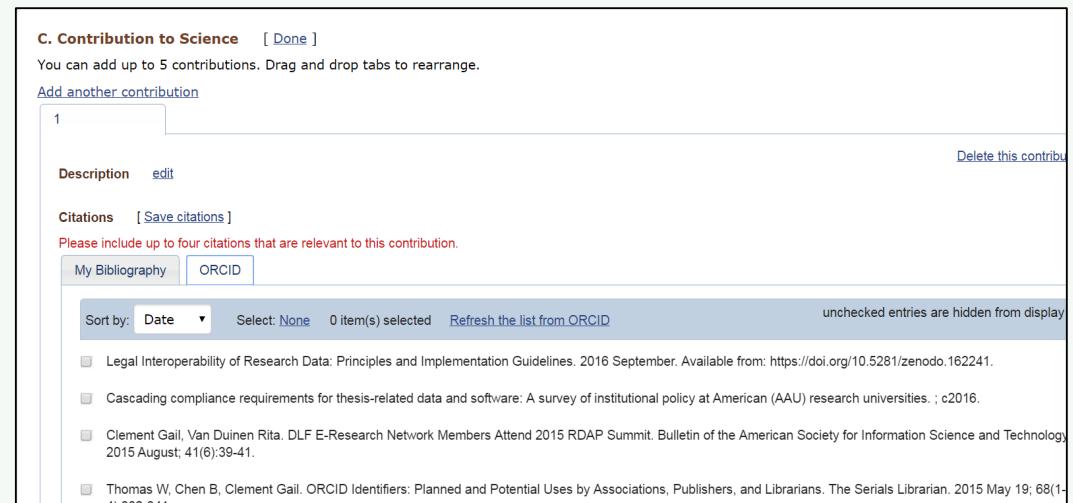
with
AuthorCarpentry

Many of today's web-based Research Information Systems rely on Your works + Your bio as linked data

U.S. Funding Agencies – SciENcv
Science Experts Network Curriculum Vitae
<https://www.ncbi.nlm.nih.gov/sciencv/>



The screenshot shows a researcher profile for Clement, Gail P. It includes a photo, contact information (Primary Email: gpcerets@caltech.edu), research interests (e.g., Legal Interoperability of Research Data, Cascading compliance requirements for thesis-related data and software), and publications (12 in the last 10 full years).

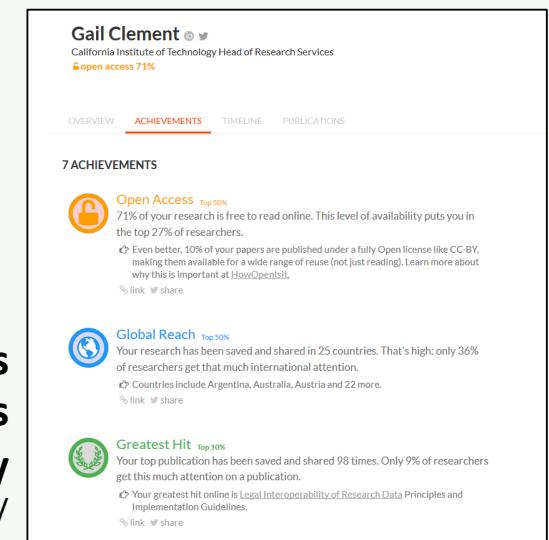


The screenshot shows a "Contribution to Science" section where you can add up to 5 contributions. It includes fields for Description, edit, and Delete this contribution. Below it is a "Citations" section with a "Save citations" button. A list of publications is shown, including:

- Legal Interoperability of Research Data: Principles and Implementation Guidelines. 2016 September. Available from: <https://doi.org/10.5281/zenodo.162241>.
- Cascading compliance requirements for thesis-related data and software: A survey of institutional policy at American (AAU) research universities. ; c2016.
- Clement Gail, Van Duinen Rita. DLF E-Research Network Members Attend 2015 RDAP Summit. Bulletin of the American Society for Information Science and Technology 2015 August; 41(6):39-41.
- Thomas W. Chen B, Clement Gail. ORCID Identifiers: Planned and Potential Uses by Associations, Publishers, and Librarians. The Serials Librarian. 2015 May 19; 68(1-4):332-341.

**Research Institutions or Universities
Researcher Profile Systems,
e.g., VIVO**
<http://vivoweb.org/>

**Scholarly Sharing repositories
and Researcher Networking Sites
e.g., ImpactStory**
<https://impactstory.org/>



The screenshot shows a researcher profile for Gail Clement. It includes sections for Overview, Achievements, Timeline, and Publications. Key achievements include:

- Open Access**: Top 50% (71% of research is free to read online)
- Global Reach**: Top 50% (research saved and shared in 25 countries)
- Greatest Hit**: Top 10% (publications saved and shared 98 times)

Citation

Surrogate of the work



ORCiD

Creator of work

DOI

Locator of the work



Learning how to
leverage
these 3 open
standards
has a **YUGE r.o.i. !**

This 5-step lesson teaches how
to apply these 3 standards to
establish & maintain your
scholarly identity and
reputation with

Efficiency * Trust * Openness *
Sustainability

5 Steps in a Nutshell

1. Start with the citation to a work you have created and link it to a persistent Web-resolvable unique identifier (DOI)
2. Establish your unique identity and authoritative profile with a persistent Web-resolvable unique identifier (ORCiD)
3. Link your works with their DOI's to your ORCiD profile – manually
4. Link your works with their DOI's to your ORCiD profile – automagically
5. Link your ORCiD profile with a scholarly impact service to generate metrics of use and attention

Step 1. Overview

Represent your work as an **open citation** with a **digital object identifier (DOI)** so it is machine readable and actionable



Copyright and Publication Status of Pre-1978 Dissertations: A Content Analysis Approach
Gail Clement and Melissa Levine

abstract: We investigated whether American dissertations that were deposited in university libraries or disseminated on microfilm prior to 1978 were "published" for copyright purposes. This question has direct bearing on the copyright status of these works today. In the absence of a directly relevant legal decision to clarify the matter, the authors examined how the former community of practice interpreted the law in the context of dissertation dissemination. A content analysis of written communications by members of this community indicates that both forms of dissertation dissemination were considered to be legal publication under the 1909 Copyright Act.

Introduction
In this study, we examine the publication and copyright status of American dissertations produced before 1978. Our particular focus is the subset of dissertations that were not distributed through formal channels such as books and journals. Rather, their only means of public access was at typescript copies shelved in the library or as microfilm reproductions available from a third-party distributor. We are interested in this particular set of dissertations because of its potential to yield public domain works, not subject to copyright restrictions and thus suitable for digital access and other scholarly uses. With an estimated 520,000 works in this category, the possibility that some portion could be made available for free public access with relatively little administrative ease would be a significant boon for digital library development and scholarship across the disciplines.

Why would some mid-20th century dissertations be in the public domain? The answer comes from an understanding of earlier copyright law enacted in 1909 and

portal: Libraries and the Academy, Vol. 11, No. 3 (2011), pp. 813-829.
Copyright ©2011 by The Johns Hopkins University Press, Baltimore, MD 21288.

Browse > Library Science and Publishing > Library and Information Science

 Copyright and Publication Status of Pre-1978 Dissertations: A Content Analysis Approach
Gail Clement, Melissa Levine

Abstract
Abstract:
We investigated whether American dissertations that were deposited in university libraries or disseminated on microfilm prior to 1978 were "published" for copyright purposes. This question has direct bearing on the copyright status of these works today. In the absence of a directly relevant legal decision to clarify the matter, the authors examined how the former community of practice interpreted the law in the context of dissertation dissemination. A content analysis of written communications by members of this community indicates that both forms of dissertation dissemination were considered to be legal publication under the 1909 Copyright Act.

From: portal: Libraries and the Academy
Volume 11, Number 3, July 2011
pp. 813-829 | 10.1353/pla.2011.0031


Open Citation 101

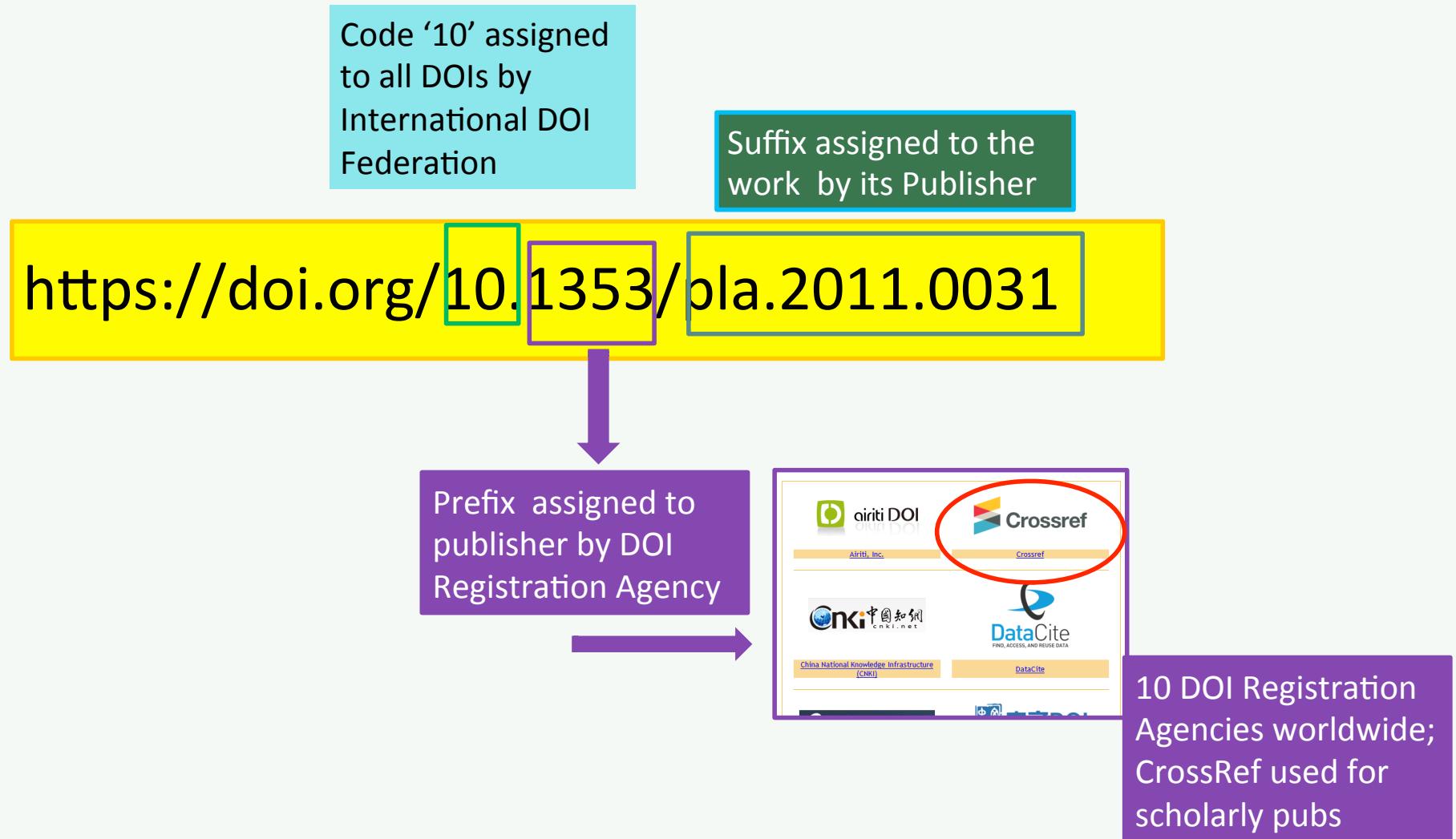
scholarly citation data which provides -- in an openly licensed, structured format -- accurate citation information (bibliographic references) harvested from the scholarly literature

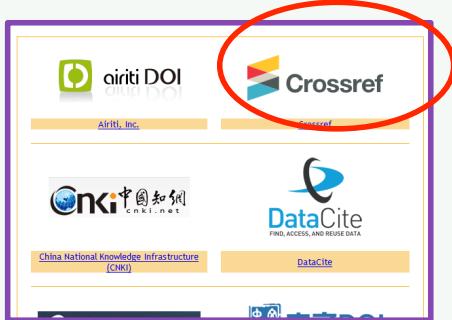
Adapted from the Open Citations Project
<http://opencitations.net/>

```
@article{Clement_2011,  
title={Copyright and Publication  
Status of Pre-1978 Dissertations:  
A Content Analysis Approach},  
volume={11}, ISSN={1530-7131},  
url={http://dx.doi.org/10.1353/  
pla.2011.0031}, DOI={10.1353/  
pla.2011.0031}, number={3},  
journal={portal: Libraries and the  
Academy}, publisher={Johns  
Hopkins University  
Press}, author={Clement, Gail  
and Levine, Melissa},  
year={2011}, pages={813–829}}
```

DOI 101

A DOI string will always resolve to a web page containing information about the work





CrossRef 101

<https://www.crossref.org/>

- One of 10 DOI Registration Agencies worldwide
- Most published articles and proceedings get their DOIs from CrossRef
- Conform to all DOI Federation requirements for assigning DOI's ***plus***
- They apply additional requirements for metadata deposit and types of works eligible and these can be **picky**
- They offer free services to find, cite, link and assess scholarly works

Step 1 to do

```
$ curl -LH "Accept: text/bibliography; style=bibtex"  
https://doi.org/10.1353/pla.2011.0031 >  
clement2011.bib
```

Step 1 to do

```
$ cat clement2011.bib
```

```
@article{Clement_2011, title={Copyright and  
Publication Status of Pre-1978 Dissertations: A Content  
Analysis Approach}, volume={11}, ISSN={1530-7131},  
url={http://dx.doi.org/10.1353/pla.2011.0031},  
DOI={10.1353/pla.2011.0031}, number={3},  
journal={portal: Libraries and the Academy},  
publisher={Johns Hopkins University  
Press}, author={Clement, Gail and Levine, Melissa},  
year={2011}, pages={813–829}}
```

Step 2. Overview

Represent your identity as a unique number using ORCID so it is machine readable and actionable

The screenshot shows the ORCID profile page for Yolanda Gil. The top navigation bar includes links for Rossman Publications, Caltech Library's Digit, Regex Tester – Regex!, Girl Develop It - Intro, Learn Enough Comm, repl.it - Compiler, Imported From IE, and a search bar. The main menu has sections for FOR RESEARCHERS, FOR ORGANIZATIONS, ABOUT, HELP, and SIGN IN. The profile page displays basic information like ORCID ID (orcid.org/0000-0001-8465-8341), Country (United States), Keywords (Artificial Intelligence), and Websites (Professional web site). Two expandable sections show her Education (Carnegie Mellon University) and Employment (University of Southern California) details.

Yolanda Gil

ORCID ID
orcid.org/0000-0001-8465-8341

Country
United States

Keywords
Artificial Intelligence

Websites
Professional web site

Education (1)

Carnegie Mellon University: Pittsburgh, PA, United States
1992
PhD (Department of Computer Science)
Source: Yolanda Gil
Created: 2015-06-04

Employment (1)

University of Southern California: Los Angeles, CA, United States
2009 to present
Research Professor (Information Sciences Institute)
Source: Yolanda Gil
Created: 2015-06-04



ORCiD 101

- Open Researcher and Contributor ID (ORCID) is three things:
(1) A membership organization; (2) a standard; (3) a profile system
- The non-profit, membership organization to solving the long-standing name ambiguity problem in scholarly communication
- The profile system maintains a central registry of unique identifiers for individual researchers
- The system supports an open, transparent linking mechanism between ORCID and other current author identifier schemes (ISO's International Standard Name Identifier (ISNI); Thomson Reuter's ResearcherID; Elsevier's SCOPUS ID)
- The ORCID itself is a unique, 16-digit identifier expressed a URL link. i.e.
<http://orcid.org/5412-3652-8965-8745>
- The ORCID URL links to the owner's profile on the ORCID website.

ORCiD 101

**ORCiD == Online Research Contributor iD ==
Global standard that solves Author Name Ambiguity Problems**

The screenshot shows the ORCID homepage with a navigation bar at the top. The navigation bar includes links for 'FOR RESEARCHERS', 'FOR ORGANIZATIONS', 'ABOUT', 'HELP', and 'SIGN OUT'. Below the navigation bar, there are links for 'MY ORCID RECORD', 'INBOX (8)', 'ACCOUNT SETTINGS', 'DEVELOPER TOOLS', and 'LEARN MORE'. A statistic '3,006,370 ORCID iDs and counting. See more...' is displayed. The main content area is titled 'Search results' and lists two entries:

ORCID iD	First name	Last name	Other names
0000-0003-3144-1645	Carol	Finn	
0000-0002-6178-0405	Carol	Finn	

ORCID 101

ORCID
Connecting Research and Researchers

FOR RESEARCHERS FOR ORGANIZATIONS ABOUT HELP SIGN IN

SIGN IN REGISTER FOR AN ORCID ID LEARN MORE

3,006,370 ORCID IDs and counting. See more...

Carol Finn

ORCID ID
 orcid.org/0000-0003-3144-1645

Country
United States

Keywords
geomagnetism, magnetic observatory

Websites
<http://geomag.usgs.gov>

Education (2)

St. Louis University: St. Louis, Missouri, United States
1986-05
Master of Professional Geophysics (Earth and Atmospheric Sciences)
Source: Carol Finn Created: 2015-09-29

Southwest Missouri State University: Springfield, Missouri, United States
1982-07
Bachelor of Science in Geology (Geology and Geography)
Source: Carol Finn Created: 2015-09-29

Employment (1)

U.S. Geological Survey: Golden, Colorado, United States
2006-10 to present
Geomagnetism Group Leader (Geologic Hazards Science Center)
Source: Carol Finn Created: 2015-09-29

Works (8)

The Boulder magnetic observatory
Open-File Report
2015 | other
DOI: [10.3133/ofr20151125](https://doi.org/10.3133/ofr20151125)
Source: CrossRef Metadata Search Preferred source

Improved Geomagnetic Referencing in the Arctic Environment
SPE Arctic and Extreme Environments Technical Conference and Exhibition
2013 | conference-paper
DOI: [10.2118/166850-ms](https://doi.org/10.2118/166850-ms)
Source: CrossRef Metadata Search Preferred source

ORCID
Connecting Research and Researchers

FOR RESEARCHERS FOR ORGANIZATIONS ABOUT HELP SIGN IN

SIGN IN REGISTER FOR AN ORCID ID LEARN MORE

3,006,370 ORCID IDs and counting. See more...

Carol Finn

ORCID ID
 orcid.org/0000-0002-6178-0405

Education (3)

University of Colorado Boulder: Boulder, CO, United States
MS (Geology)
Source: Carol Finn Created: 2014-05-22

University of Colorado Boulder: Boulder, CO, United States
Ph.D. (Geology)
Source: Carol Finn Created: 2014-05-22

Wellesley College: Wellesley, MA, United States
BA (Geology)
Source: Carol Finn Created: 2014-05-22

Employment (1)

US Geological Survey: Denver, CO, United States
Research scientist (Crustal Geophysics and Geochemistry)
Source: Carol Finn Created: 2014-05-22

Works (1)

Mapping the 3D extent of the Northern Lobe of the Bushveld layered mafic intrusion from geophysical data
Precambrian Research
2015-10 | journal-article
DOI: [10.1016/j.precamres.2015.07.003](https://doi.org/10.1016/j.precamres.2015.07.003)
Source: Crossref Preferred source

More and More Publishers Requiring ORCID iDs

American Chemical Society

eLife

PLOS

The Royal Society

American Geophysical Union & Wiley generally

EMBO Press

Hindawi

IEEE

Science

INFORMS

Faculty of 1000

Wellcome Open Research

Rockefeller University Press

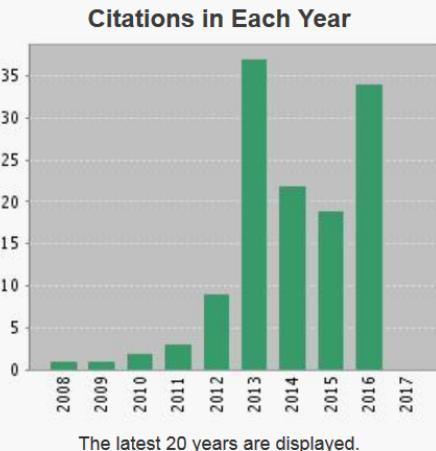
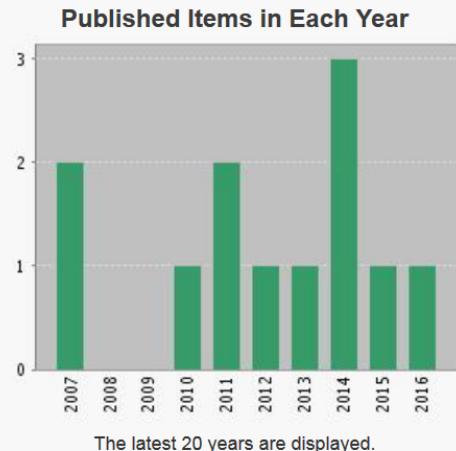
Author Name Ambiguity Problems

Citation Report: 12

(from Web of Science Core Collection)

You searched for: AUTHOR: (finn carol) [...More](#)

This report reflects citations to source items indexed within Web of Science Core Collection. Perform a Cited Reference Search to include citations to items not indexed within Web of Science Core Collection.



Results found: 12

Sum of the Times Cited [\[?\]](#) : 128

Sum of Times Cited without self-citations [\[?\]](#) : 122

Citing Articles [\[?\]](#) : 118

Citing Articles without self-citations [\[?\]](#) : 113

Average Citations per Item [\[?\]](#) : 10.67

h-index [\[?\]](#) : 4

Is this Author getting credit where due?

Is this Author self-citing her own work?

Is this Author accruing accurate citation metrics?

Is this Author editing or reviewing her own papers?

Is this researcher double-dipping across funding agencies?

Step 2 to do

Point your browser to
<https://orcid.org/>

Click on Register now!
to sign up with your
email

Check your email and
click through the link
provided

DISTINGUISH YOURSELF IN THREE EASY STEPS

ORCID provides a persistent digital identifier that distinguishes you from every other researcher and, through integration in key research workflows such as manuscript and grant submission, supports automated linkages between you and your professional activities ensuring that your work is recognized. [Find out more](#).



REGISTER

Get your unique ORCID identifier [Register now!](#)
Registration takes 30 seconds.

Steps 3 + 4 Overview

Connect your citations with DOIs to your ORCID works

The screenshot shows an ORCID profile for Gail P. Clement. On the left, there's a sidebar with sections for 'Also known as', 'Country' (United States), 'Keywords' (Authorship and Attribution ethics, Copyright education, Scholarly communication, scientific publishing, Research data dissemination and publication), 'Websites' (Ask A Librarian at Caltech), 'Email' (gclement@library.caltech.edu, gperetsm@caltech.edu), and 'Other IDs' (Scopus Author ID: 43461134500, Scopus Author ID: 7102259798, ISNI: 0000000137477695, ISNI: 000000063976212). The main area displays three sections: 'Education (1)', 'Employment (1)', and 'Works (26)'. The 'Works' section is circled in red. It contains four items:

- Legal Interoperability of Research Data: Principles and Implementation Guidelines (2016-09-08 | report, URL: <https://doi.org/10.5281/zenodo.162241>, Source: Gail P. Clement, Preferred source)
- Using the Research of Others Responsibly: Caltech Library (2016-05-27 | lecture-speech, DOI: [10.7907/Z97H1GJW](https://doi.org/10.7907/Z97H1GJW), URL: <http://resolver.caltech.edu/CaltechAUTHORS:20160526-082952104>, Source: Gail P. Clement, Preferred source)
- 21st Century Scientific Authoring at Institutions: Demonstrating the Overleaf Caltech Portal (Figshare, 2016 | data-set, DOI: [10.6084/M9.FIGSHARE.3171994](https://doi.org/10.6084/M9.FIGSHARE.3171994), Source: DataCite, Preferred source)
- ORCID Identifiers: Planned and Potential Uses by Associations, Publishers, and Librarians (The Serials Librarian, 2015-05 | journal-article, Source: DataCite, Preferred source)

Step 3 to do

Connect your citations with DOIs to your ORCID works, manually

The screenshot shows the ORCID profile page under the 'Works' tab, which contains 29 items. A yellow sticky note labeled 'mypub.bib' is overlaid on the bottom left. On the right, a red circle highlights the 'Import BibTeX' button, with a red arrow pointing from the note towards it. The 'Import BibTeX' button is located in a vertical menu next to the 'Add works' button.

▼ Works (29)

21st Century Scientific Authoring at Institutions 30s Mov
Figshare
2016 | other
DOI: [10.6084/M9.FIGSHARE.3174229](https://doi.org/10.6084/M9.FIGSHARE.3174229)

Source: DataCite

[Import BibTeX](#)

+ Add works

Search & link

Import BibTeX

+ Add manually

mypub.bib

Step 4 to do

Connect your citations with DOIs to your ORCID works, auto-magically

Source: Gail P. Clement Created: 2013-12-09  

Funding (0)  Add funding 

You haven't added any funding, [add some now](#)

Works (28)  Add works  Bulk edit 

LINK WORKS [Hide link works](#)

ORCID works with our member organizations to make it easy to connect your ORCID iD and link to information in their records. Choose one of the link wizards to get started. [More information about linking works](#)

Airiti
Enables user to import metadata from Airiti, including journal papers, proceedings, dissertations ... 

Australian National Data Service (ANDS) Registry
Import your research datasets into ORCID from Australian National Data Service (ANDS) and Res... 

CrossRef Metadata Search
Import your publications from CrossRef's authoritative, publisher-supplied metadata on over 70 ... 

Step 4 to do

 Gail P. Clement
<http://orcid.org/0000-0001-5494-4806> ▾
[\(Not You?\)](#)

CrossRef Metadata Search  has asked for the following access to your ORCID Record

Add works
Read your ORCID record

This application will not be able to see your ORCID password, or other private info in your ORCID Record. [Privacy Policy](#).

[Deny](#) [Authorize](#)

Step 4 to do

Crossref

TYPE Journal Article (2,346)

YEAR 2011 (2,346)

PUBLICATION

- PLoS ONE (41)
- Physics Letters B (36)
- Journal of High Energy Physics (35)
- Physical Review Letters (34)
- Cancer Research (17)
- The European Physical Journal C (17)
- Physical Review D (15)
- ChemInform (11)
- Genetic Engineering & Biotechnology News (11)

Gail Clement

Status API Help  Gail Clement ▾

SORT BY: RELEVANCE PUBLICATION YEAR

PAGE 1 OF 2,346 RESULTS

Copyright and Publication Status of Pre-1978 Dissertations: A Content Analysis Approach

Journal Article published 2011 in portal: Libraries and the Academy volume 11 issue 3 on pages 813 to 829

Authors: Gail Clement, Melissa Levine

<https://doi.org/10.1353/pla.2011.0031> [Actions](#)  IN YOUR PROFILE

The Basis of Differential Responses to Folic Acid Supplementation

Journal Article published 2011 in Journal of Nutrigenetics and Nutrigenomics volume 4 issue 2 on pages 99 to 109

Authors: Ioana Cotlarcic, Toby Andrew, Tracy Dew, Gail Clement, Raj Gill, Gabriela Surdulescu, Roy Sherwood, Kourosh R. Ahmadi

<https://doi.org/10.1159/000327768> [Actions](#)  ADD TO ORCID

Step 4 to do

Manage duplicate entries in your ORCID works

Copyright and Publication Status of Pre-1978 Dissertations: A Content Analysis Approach
portal: Libraries and the Academy
2011 | journal-article
DOI: [10.1353/pla.2011.0031](https://doi.org/10.1353/pla.2011.0031)
URL: <http://dx.doi.org/10.1353/pla.2011.0031>

URL
<http://dx.doi.org/10.1353/pla.2011.0031>

Citation (bibtex) [switch view]
@article{Clement_2011, title= {Copyright and Publication Status of Pre-1978 Dissertations: A Content Analysis Approach}, volume= {11}, ISSN= {1530-7131}, url= {http://dx.doi.org /10.1353/pla.2011.0031}, DOI= {10.1353/pla.2011.0031}, number= {3}, journal= {portal: Libraries and the Academy}, publisher= {Johns Hopkins University Press}, author= {Clement, Gail and Levine, Melissa}, year= {2011}, pages= {813-829}}

Created
2017-01-23

Source: Gail P. Clement

Preferred source (of 2)

Step 5 Overview

Connect your citations, DOI's and ORCiD profile to a scholarly impact service to generate metrics

The screenshot shows a researcher's profile on Impactstory. At the top, there is a photo of Ethan White, his name, and his affiliation with the University of Florida. Below this, a black banner reads "Alternative Metrics: Measures of attention and usage beyond the Impact Factor". A large red oval highlights the "TIMELINE" section, which displays a total of 3867 online mentions over 1 year, with specific counts for various platforms: 3.6k (RSS), 100 (DOI), 97 (Google Scholar), 32 (CrossRef), 8 (Publons), 7 (Scopus), 6 (PubMed), 6 (Dimensions), and 4 (Altmetric). A blue arrow points down from the banner towards the timeline section. To the left of the timeline, there is a "ACHIEVEMENTS" section with three items: "Wikitastic Top 10%" (mentioned in 6 Wikipedia articles), "Open Access Top 25%" (86% free online), and "Hot Streak Top 10%" (research talked about monthly). Below the timeline is a "PUBLICATIONS" section listing three papers: "Best Practices for Scientific Computing" (2014 PLoS Biology), "The Case for Open Preprints in Biology" (2013), and "Elevating The Status of Code in Ecology" (2014). At the bottom, the Impactstory logo is shown with the tagline "Discover the online impact of your research: Join for free with Twitter".



Impactstory

<https://www.impactstory.org/>

ImpactStory 101

- A non-profit web service by and for researchers
- Provides alternative measures of attention and usage for scholarly works (“Altmetrics”)
- Awarded a \$297,500 EAGER grant from the National Science Foundation to study how automatically-gathered impact metrics can improve the reuse of research software.
- Integrates with ORCID for citation and author data 
- Integrates with Elsevier/SCOPUS for citation data 

Step 5 Overview



Alternative Metrics (Altmetrics): Measures of attention and usage for scholarly works beyond the Impact Factor or H-Index

[◀ Back to Ethan's publications](#)

Best Practices for Scientific Computing

Wilson, G., Aruliah, D.A., Brown, C.T., Chue Hong, N.P., Davis, M., Guy, R.T., Haddock, S.H.D., Huff, K.D., Mitchell, I.M., Plumley, M.D. et al.
2014 PLoS Biology ↗

🔒 Free fulltext available ↗

SHARED 5 TIMES ON news ×

- 📰 Scientific computing: Code alert
7 days ago by Nature
📄 Best Practices for Scientific Computing
- 📰 How Computers Broke Science—And What We Can Do to Fix It
a year ago by The Epoch Times
📄 Best Practices for Scientific Computing
- 📰 How Computers Broke Science—and What We Can Do to Fix It
a year ago by Gizmodo
📄 Best Practices for Scientific Computing
- 📰 How Computers Broke Science | RealClearScience
a year ago by Real Clear Science
🔗 DOI: 10.1371/journal.pbio.1002001

Discover the online impact of your research: [Join for free with Twitter](#)

Filter by activity

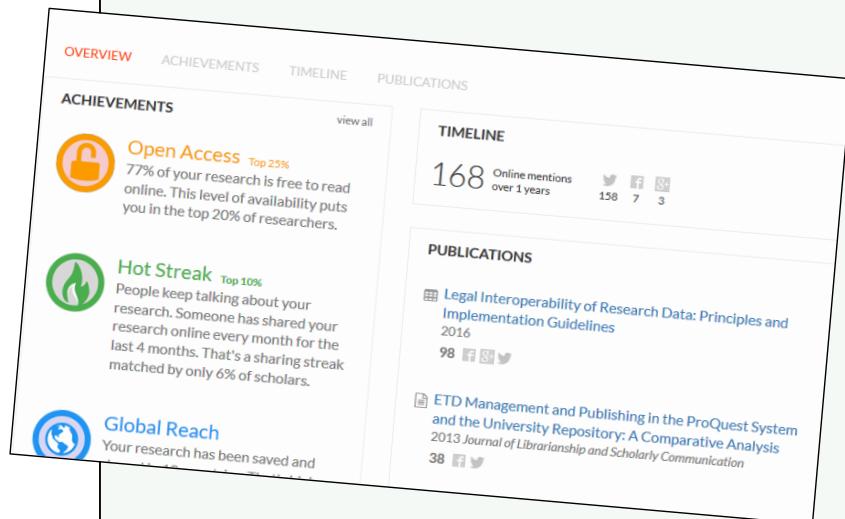
- Tweets (2k ↑)
- Google+ posts (70)
- Blog posts (35)
- Facebook pages (15)
- Reddit posts (7)
- ×
- News mentions (5 ↑)
 - Q&A post mentions (4)
 - Weibo posts (1)
 - Wikipedia articles (1)

Step 5 to do

Point your browser to
<https://impactstory.org>

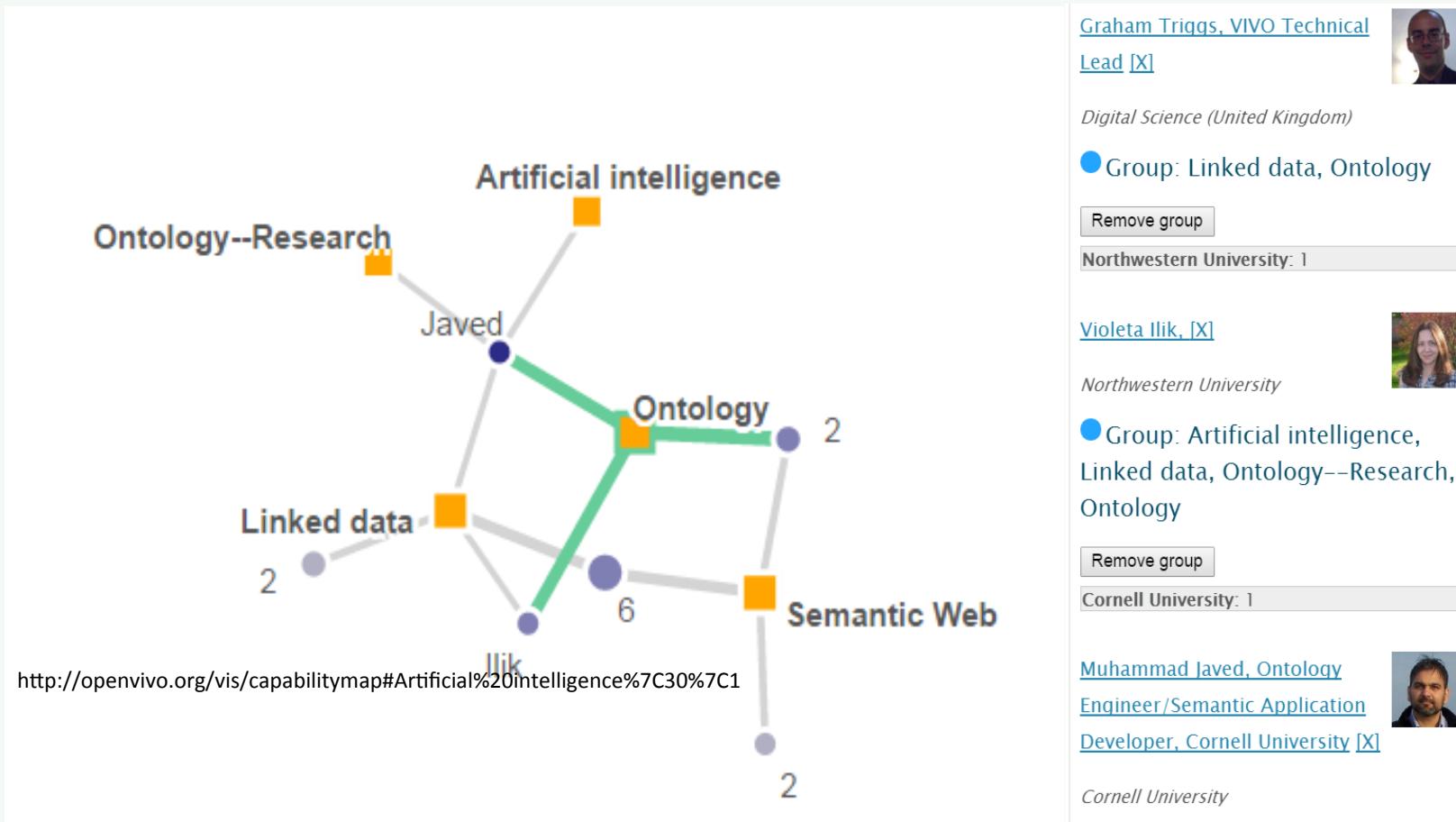
Click Login to sign up with
your ORCiD

Voila! Watch your impact
profile auto-populate thanks
to linkages between those 3
standards (citation, DOI,
ORCiD)



Next Steps?

Explore what other linkages and networks you can generate with your 3 linked data building blocks





Index Gail

Search

Home | People | Organizations | Research | Events | Capability Map |

Welcome back, Gail Clement

OpenVIVO has found works for you to claim. For each work, you will have an opportunity to indicate your role in the work.

There are 3 works to claim.

Click the "Claim these works" to claim the works. Click Cancel if you do not wish to claim works at this time.

[Claim these works](#)

[Cancel](#)

21st Century Scientific Authoring at Institutions: Demonstrating the Overleaf Caltech Portal

2016;

Miller, S;  Clement, G

Unlisted Author

Editor

Other Contribution

Please indicate the contribution that you made to this work

[Hide roles](#)

Author

Writing Original Draft

Figure Development

Editing and Proofreading

Translator

Background and Literature Search

Conceptualization

Preservation

Archivist

Digital Preservation

Data

Data Curation

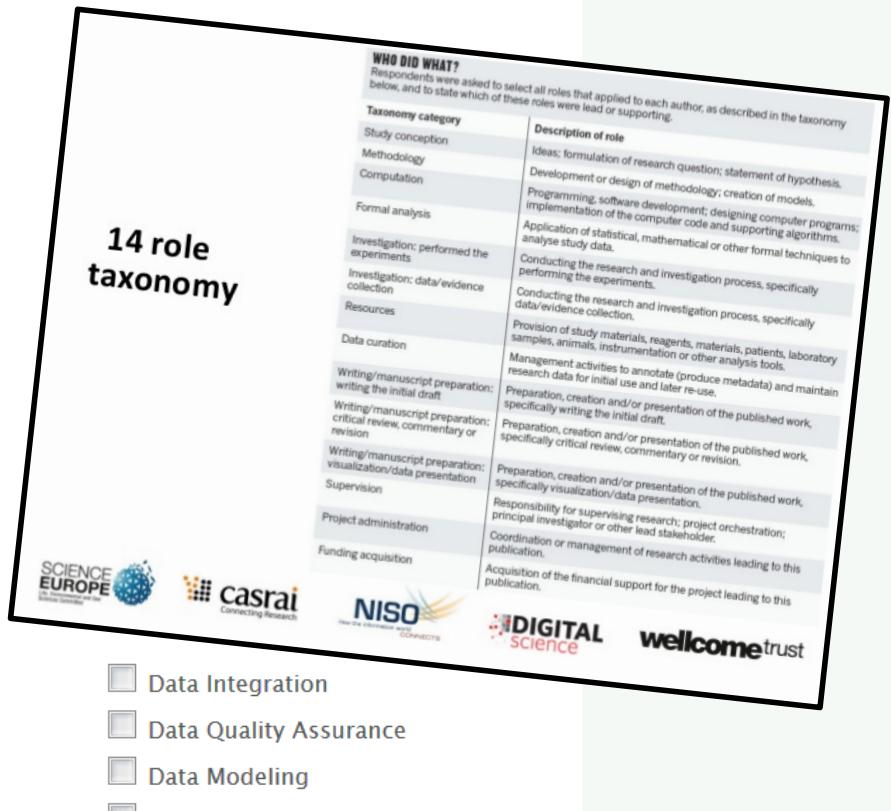
Data Analysis

Metadata Application

Statistical Data Analysis

Data Entry

Data Collection



VIVO applies the Contributor Role taxonomy developed through Project CRediT

Funding Acquisition

- Study Investigation

Methodology

- Technique Development
- Protocol Creation
- Project Management
- Team Management
- Regulatory Administration
- Policy Development

- Guideline Development

- Study Design

- Standard Operating Procedure Development

Communication

- Marketing
- Networking Facilitation

- Graphic Design
- Website Development

- Documentation

Software Developer

- Software Architecture
- Software Design
- Computer Programming

- Software Engineering
- Software Testing
- Software Project Management

- Code Review
- Technical Writing

Information Technology Systems

- Software Systems
- Supervision
- Validation

- Database Administrator

- Hardware Systems

Research Instrumentation

- Device Development

- Equipment Technician

- Survey and Questionnaire

Educational

VIVO applies the Contributor Role taxonomy
developed through Project CRediT

Recommended Resources



AuthorCarpentry | <https://authorcarpentry.github.io>

Contributor Roles Taxonomy (Project CRediT) | <http://docs.casrai.org/CRediT>

CrossRef DOI Registration Agency | <https://corssref.org/>

CrossRef API Documentation | https://github.com/CrossRef/rest-api-doc/blob/master/rest_api.md

DOI Federation | <https://www.doi.org/>

ImpactStory | <https://impactstory.org>

Laure Haak's ORCID Blog – Publishers Starting to Require ORCIDs, Jan 7, 2016
<https://orcid.org/blog/2016/01/07/publishers-start-requiring-orcid-ids>

Open Citations Project, <http://opencitations.net/>

SciENcv, National Library of Medicine, <https://www.ncbi.nlm.nih.gov/sciencv/>

SciENcv: Integrating with ORCID (video demo) https://www.youtube.com/watch?v=G_ckSRr7TJ4&feature=youtu.be

“Scientists Your Number is up”. *Nature News* 485(7400) May 30 2012
<http://www.nature.com/news/scientists-your-number-is-up-1.10740>

VIVO Scholarly Networking System | <http://vivoweb.org/>

Thank you!

AuthorCarp^{entry}

The Scientific Paper of the Future: An Author Checklist

OntoSoft Training

Part 7

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

<http://www.scientificpaperofthefuture.org>



CC-BY
Attribution



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.

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)

Review of Best Practices: Author Checklist

1

Data accessibility

2

Data documentation

3

Software accessibility

4

Software documentation

5

Provenance documentation

6

Methods documentation

7

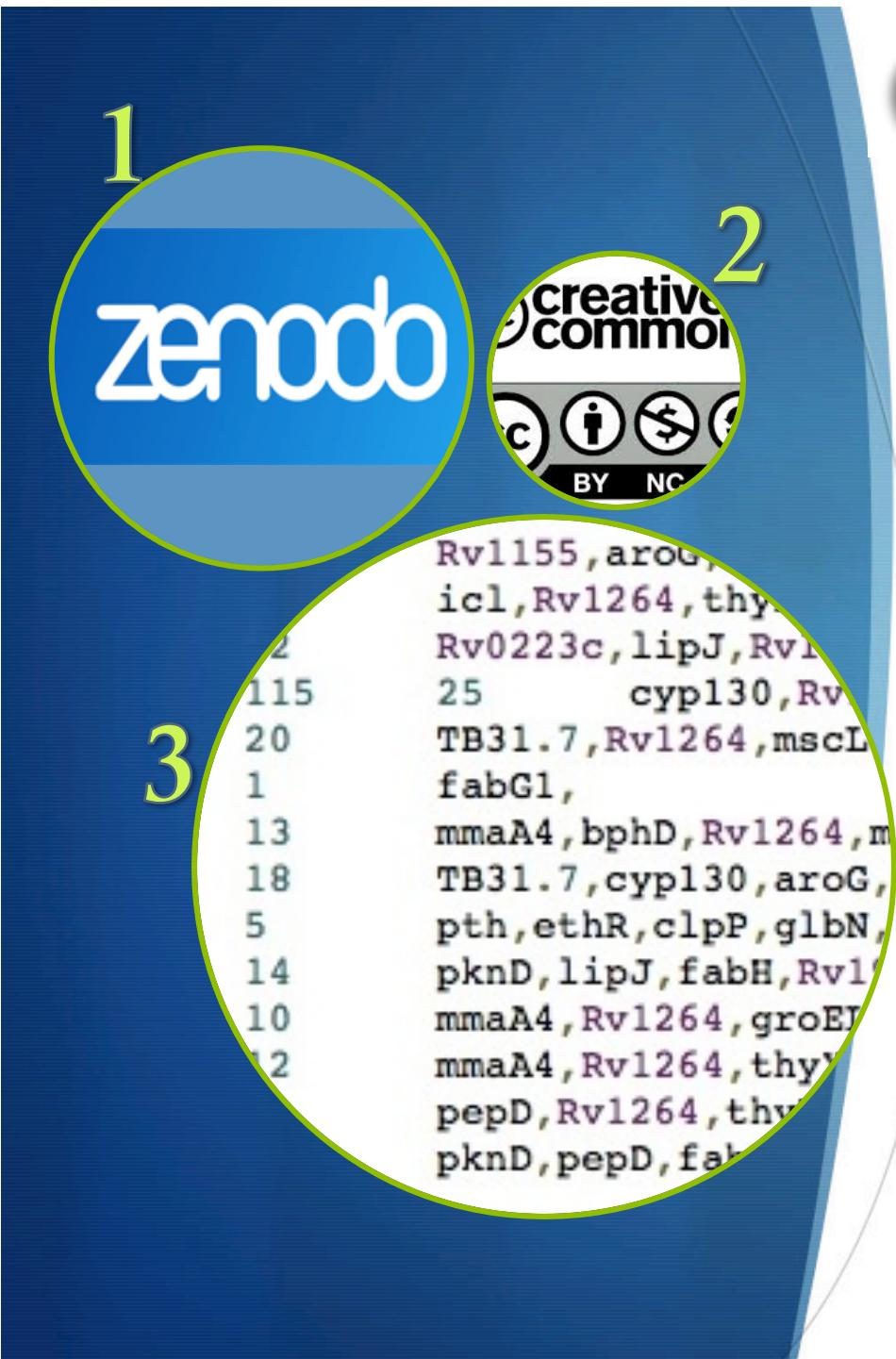
Authors identification

1

Data accessibility

Simplest Approach

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 you a data citation**



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 required by that repository using metadata standards for that community
 - Including license -- choose from <http://www.creativecommons.org/licenses>
4. Upload/point to the data
5. Get a data citation from the repository



What to Show in the Paper

- ★ 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

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

Permanent unique identifier

Name

Repository



CF MetaData

ISO 19115

WaterML2.0

2

Data documentation

Simplest Approach

- ★ Datasets should have general-purpose metadata specified (creator, date, name, etc.)

Ideal Approach

- ★ Dataset characteristics should be explained in detail
- ★ Domain-specific metadata should be documented
- ★ Availability of related datasets should be documented

What to Show in the Paper

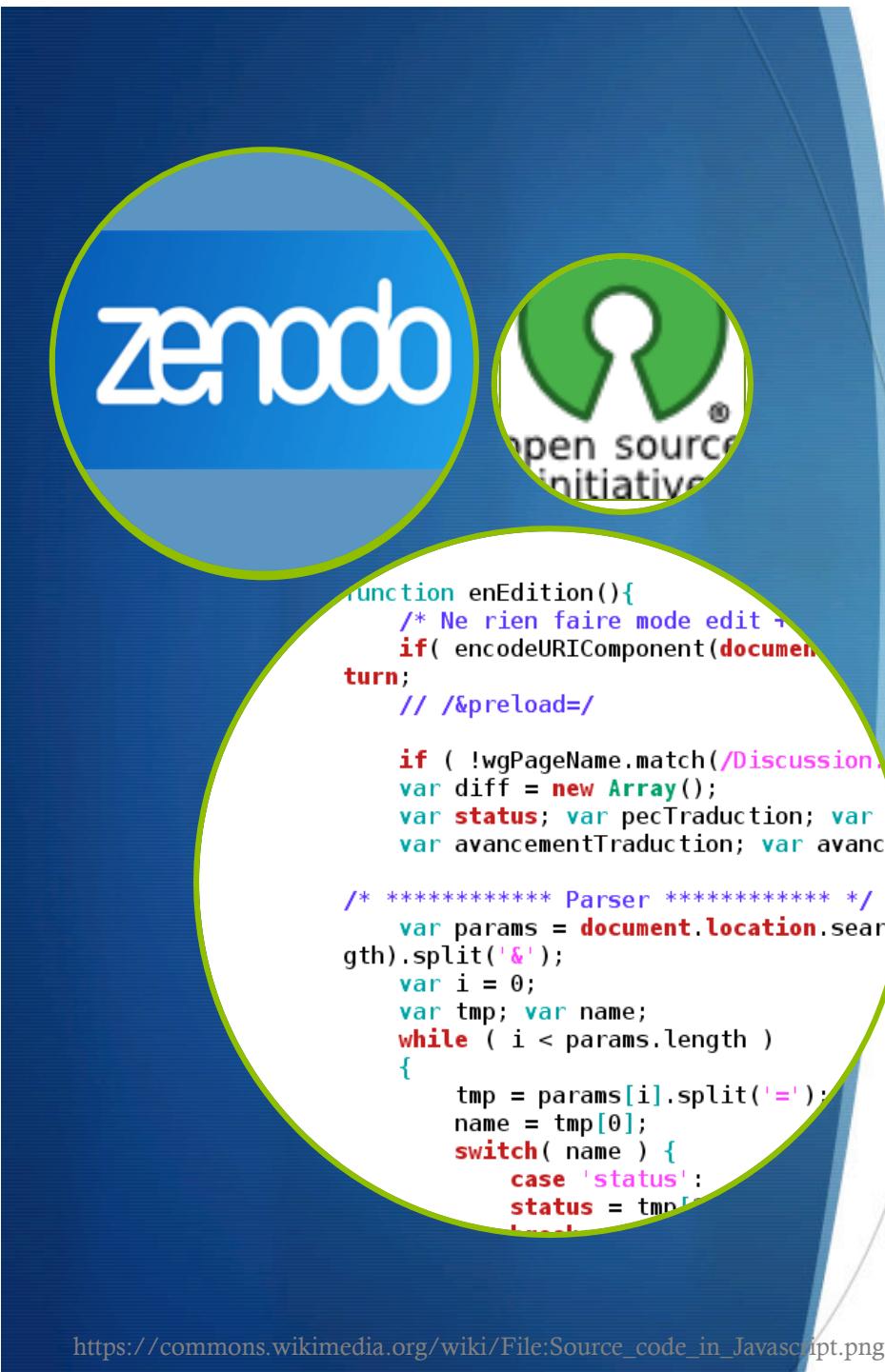
The screenshot shows a data documentation page from the LTER Network. At the top, there's a logo for LTER and the text "The US Long Term Ecological Research Network". Below that, a header bar says "WDNR Yahara Lakes Fisheries: Fish Lengths and Weights 1987-1998 - Lathrop". The main content area includes:

- LTER Identifier:** knb-lter-ntl.279.1
- Abstract:** A detailed paragraph explaining the data collection period (1987-1998), the transition to the University of Wisconsin's NTL-LTER program, and the inclusion of WDNR data from 1997-1998. It also mentions a joint project with the University of Wisconsin-Madison, Center for Limnology (CFL) on Lake Mendota.
- Owners/Creators:** Lathrop
- Metadata:** A link to "Select [here](#) for full metadata"
- Data File(s):** A list of nine CSV files:
 - 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

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



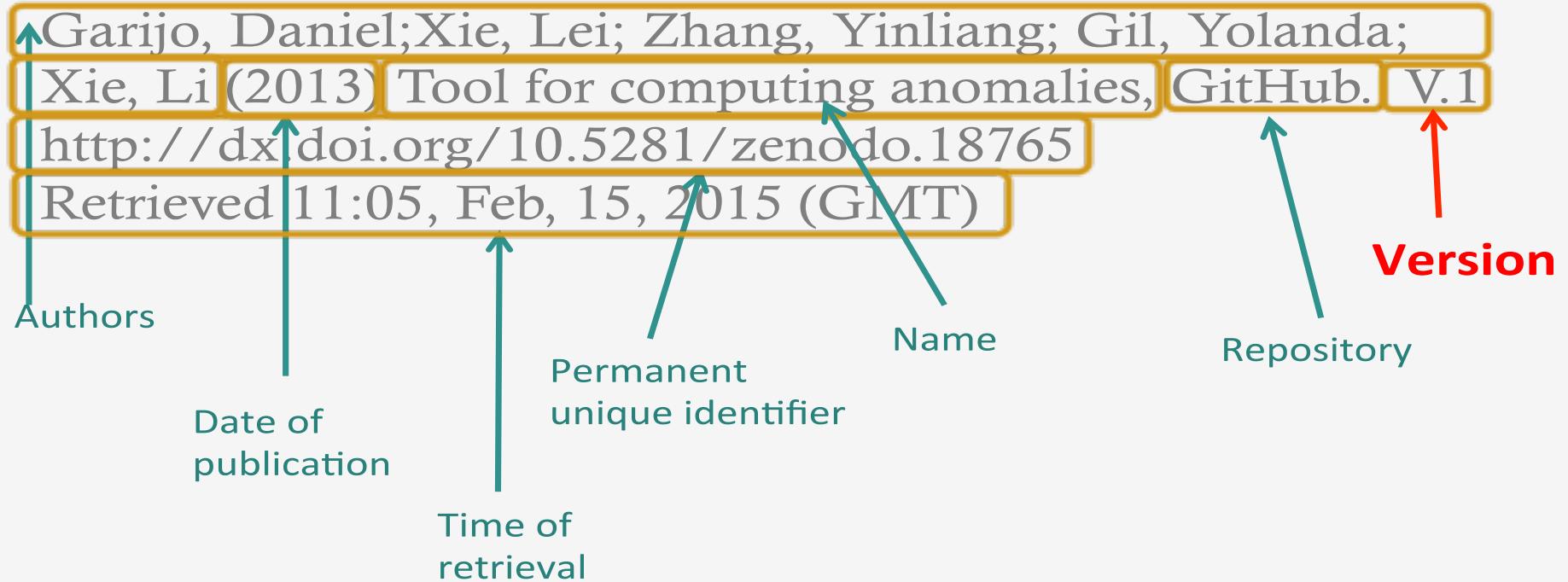
Ideal Approach

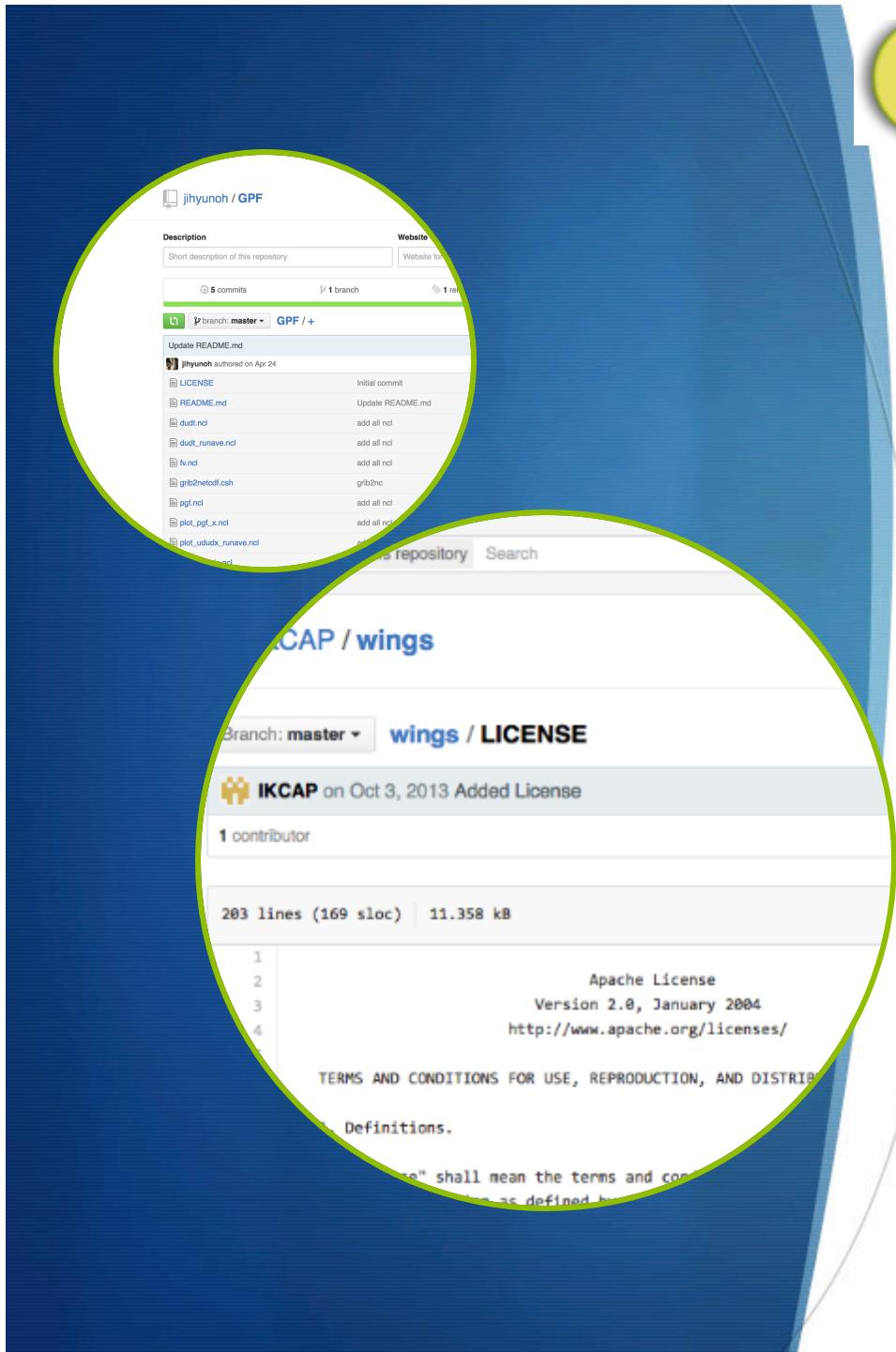
1. Learn to use a code repository that allows version tracking and collaborative software development
 - GitHub, BitBucket, etc.
2. Create a public entry for your software with a persistent unique identifier
3. Specify the metadata
 - Including license -- choose from <http://opensource.org/licenses>,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

What to Show in the Paper

- ★ 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

Software Citation Format





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

The image shows a screenshot of the OntoSoft software documentation interface. At the top left is the OntoSoft logo. Below it is a circular navigation menu divided into six segments: Identify, Understand, Execute, Do Research, Get Support, and Update. To the right of the menu is a large white circle containing sections for 'Understand' (with 'Trust - Quality and ratings'), 'Who created this software? (Project, Organization, Person, Initiative, etc.)' (listing Christopher Duffy), 'Are there any additional contributors of note for this software?' (listing Kesh Kumar Bhatt), and 'Are features of this software worth highlighting?'.

Ideal Approach

1. **Use 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**

What to Show in the Paper

- ★ 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. PIHM 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

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

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, which was scanned in order to generate a representation of the

```
CardFormatNode_7  
-----  
/usr/share/tomcat6/storage/users/admin/Water/code/library/  
/usr/share/tomcat6/storage/users/admin/Water/data/CDEC_VW...
```

```
CreateParametersFileNode_9  
-----  
/usr/share/tomcat6/storage/users/admin/Water/code/library/CreateParametersFile...  
/usr/share/tomcat6/storage/users/admin/Water/data/AvgHourly_SMN_2010-03-01...
```

```
ReaerationCMNode  
-----  
/usr/share/tomcat6/storage/users/admin/Water/code/library/ReaerationCM/run -o...  
/usr/share/tomcat6/storage/users/admin/Water/data/Params_SMN_2010-03-032  
/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/CreateParametersFile...  
/usr/share/tomcat6/storage/users/admin/Water/data/AvgHourly_SMN_2010-03-032
```

```
CreateParametersFileNode_5  
-----  
/usr/share/tomcat6/storage/users/admin/Water/code/library/CreateParametersFile...  
/usr/share/tomcat6/storage/users/admin/Water/data/AvgHourly_SMN_2010-03-032  
/usr/share/tomcat6/storage/users/admin/Water/code/library/CreateParametersFile...  
/usr/share/tomcat6/storage/users/admin/Water/data/AvgHourly_SMN_2010-03-032
```

```
CalculateHourlyAveragesNode_6  
-----  
/usr/share/tomcat6/storage/users/admin/Water/code/library/CalculateHourlyAverag...  
/usr/share/tomcat6/storage/users/admin/Water/data/AvgHourly_SMN_2010-03-032
```

5 Provenance documentation

6 Methods documentation

Simplest Approach

1. Describe the workflow in text

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

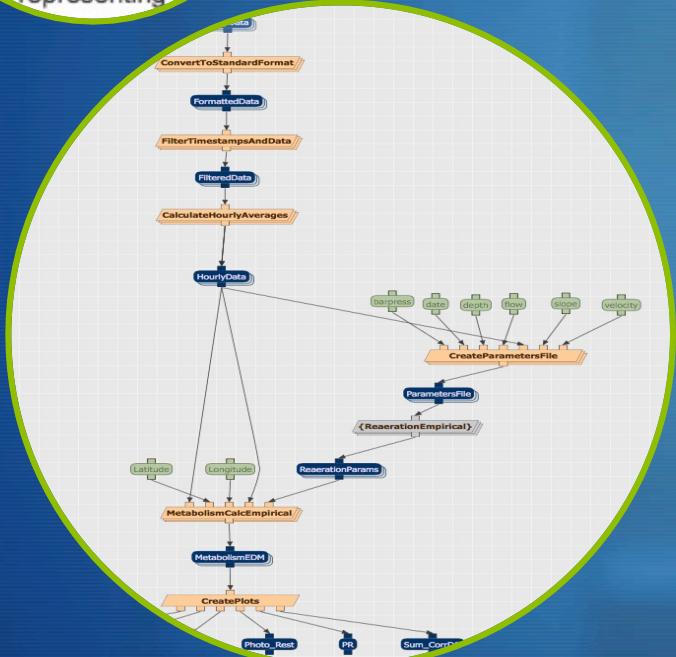
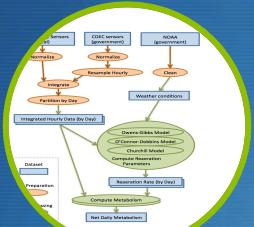
2. Develop a workflow sketch

- Capture high-level dataflow across components

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

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

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



5

Provenance documentation

6

Methods documentation

Ideal 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.

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

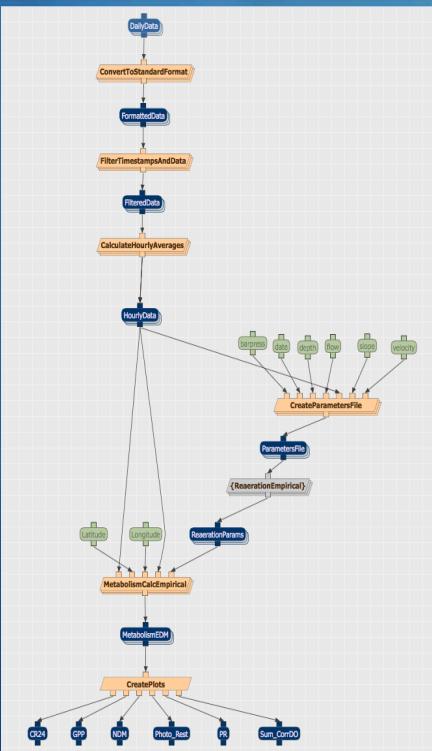
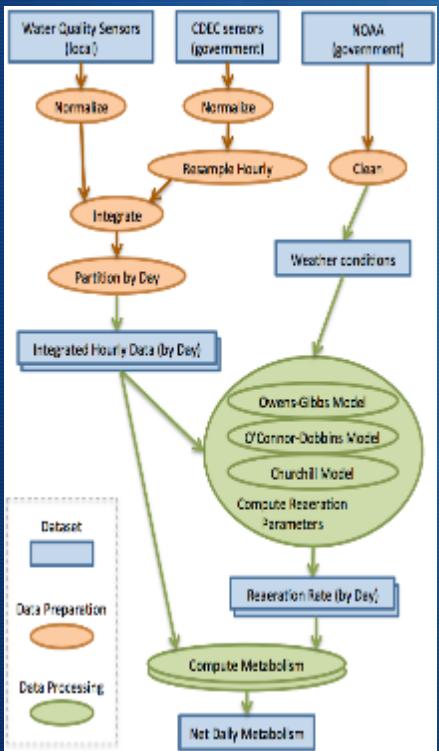
5

Provenance documentation

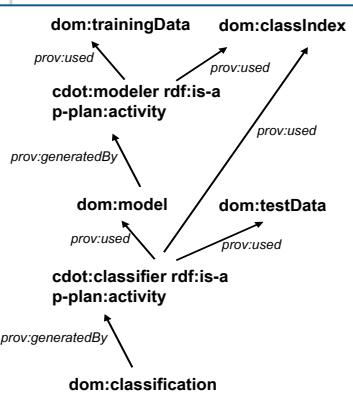
6

Methods documentation

What to Show in the Paper



- ★ 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 records using a standard like W3C PROV



```

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

# Activities
ex:Classifier1 a prov:Activity .

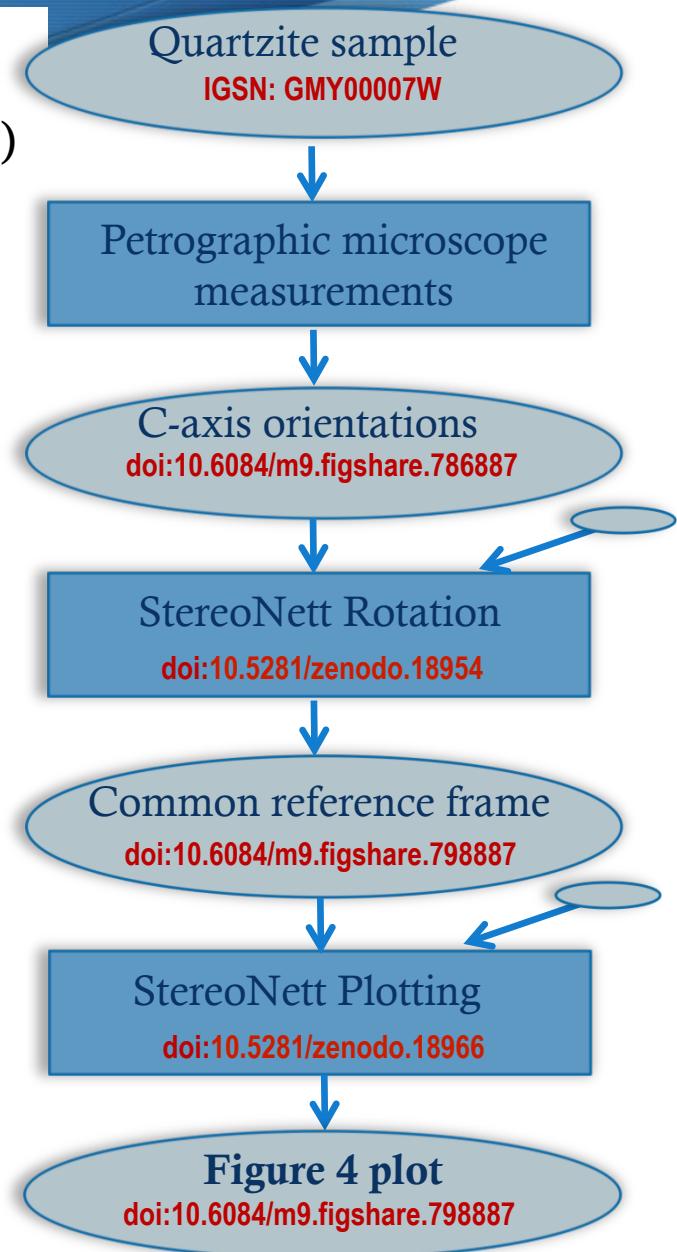
# Usage and Generation relations between entities and activities
ex:Classifier1
prov:used ex:testData1 ;
prov:used ex:model1 .
ex:classification1
prov:wasGeneratedBy ex:Classifier1 .
  
```

SPF authors should publish and cite the initial sample (or initial data), the intermediate data, the software, the final results, and the provenance of each figure or result

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

Jade Silverstein (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)), shown in Figure 4. **The provenance is shown in Fig 5.** [...]



What to Show in the Paper



- ★ Authors have a persistent unique identifier
- ★ Use www.orcid.org

ORCID



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.
- ★ **For authors**, there should be a unique identifier (e.g., ORCID)

What You Have Learned 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**



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



Incorporate Best Practices Into Your Work

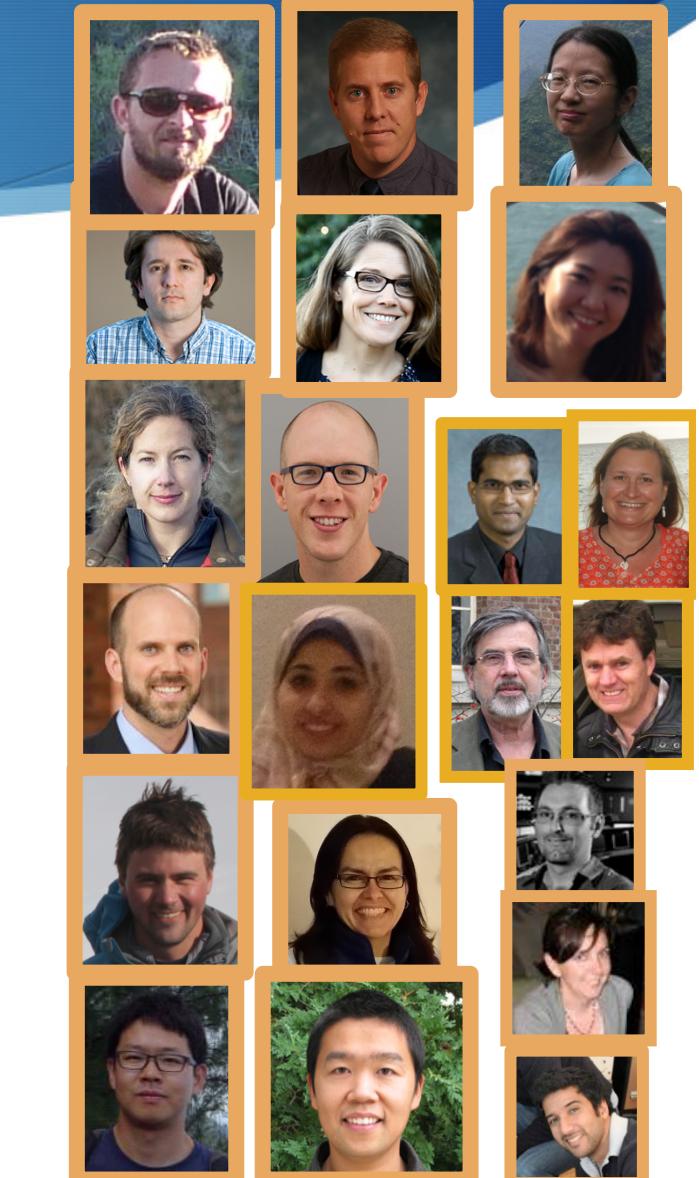
- Easier to track research products, report to funders, get credit, etc.
- Making a paper into an SPF is then very straightforward



Author Support

- ★ **Public mailing list for authors:**
<http://mailman.isi.edu/mailman/listinfo/spf-authors>
 - ★ General questions
 - ★ Approaches and tools used
 - ★ Best practices in specific disciplines

- ★ **Public mailing list for announcements:**
<http://mailman.isi.edu/mailman/listinfo/spf-announce>
 - ★ Training sessions
 - ★ Formation of author groups
 - ★ Special issues



Acknowledgments



ICER-1440323
ICER-1343800

- ★ 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 the EarthCube program

For More Information

<http://www.scientificpaperofthefuture.org>



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



[CER-1440323
[CER-1343800

