The Scientific Paper of the Future

OntoSoft Training

January 2017

ontosoft@gmail.com

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

http://www.scientificpaperofthefuture.org

EarthCube
Instructors Today

Daniel Garijo
Information Sciences Institute
University of Southern California

Yolanda Gil

Gail Clement
Research Library
California Institute of Technology
Acknowledgments

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
OntoSoft: Software Stewardship for the Geosciences

Community
- Recommender system
- Interoperability

Publication
- Structured metadata
- Interactive advice

Learning
- Best practices
- Multimedia lessons
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
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 access

Open publications

Open source
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.

<table>
<thead>
<tr>
<th>Measure of Scientific Contribution</th>
<th>No. of times chosen</th>
<th>Relative ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Publication in high-impact journals</td>
<td>92</td>
<td>2.61</td>
</tr>
<tr>
<td>Grants earned</td>
<td>65</td>
<td>1.73</td>
</tr>
<tr>
<td>Training and mentoring students and postdocs</td>
<td>63</td>
<td>1.71</td>
</tr>
<tr>
<td>No. of citations on published research</td>
<td>58</td>
<td>1.62</td>
</tr>
<tr>
<td>No. of publications</td>
<td>53</td>
<td>1.38</td>
</tr>
<tr>
<td>Teaching courses</td>
<td>41</td>
<td>1.18</td>
</tr>
<tr>
<td>Collaborative work outside of your department/institution</td>
<td>37</td>
<td>0.97</td>
</tr>
<tr>
<td>Development of research resources for the scientific community (e.g. reagents, software, database development)</td>
<td>31</td>
<td>0.89</td>
</tr>
<tr>
<td>No. of students or postdocs who go on to prestigious jobs</td>
<td>25</td>
<td>0.66</td>
</tr>
</tbody>
</table>

Other Measures:
- 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)

http://www.nature.com/nature/newspdf/metrics_survey.pur
Scientists Are Changing

[Holmberg et al 2014]

http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0106086
Publishers Are Changing

Reporting Checklist For Life Sciences Articles

1. How was the sample size chosen to ensure adequate power to detect a pre-specified effect size? For animal studies, include a statement about sample size estimate even if no statistical methods were used.

2. Describe inclusion/exclusion criteria if samples or animals were excluded from the analysis. Were the criteria pre-established?

3. If a method of randomization was used to determine how samples/animals were allocated to experimental groups and processed, describe it. For animal studies, include a statement about randomization even if no randomization was used.

4. If the investigator was blinded to the group allocation during the experiment and/or when assessing the outcome, state the extent of blinding. For animal studies, include a statement about blinding even if no blinding was done.

5. For every figure, are statistical tests justified as appropriate? Do the data meet the assumptions of the tests (e.g., normal distribution)? Is there an estimate of variation within each group of data? Is the variance similar between the groups that are being statistically compared?

Figure legends
Each figure legend should contain, for each panel where they are relevant:
- the exact sample size (n) for each experimental group/condition, given as a number, not a range;
- a description of the sample collection allowing the reader to understand whether the samples represent technical or biological replicates (including how many animals, litters, cultures, etc.);
- a statement of how many times the experiment shown was replicated in the laboratory;
- definitions of statistical methods and measures:
  - very common tests, such as t-test, simple χ² tests, Wilcoxon and Mann-Whitney tests, can be unambiguously identified by name only, but more complex techniques should be described in the methods section;
  - are tests one-sided or two-sided?
  - are there adjustments for multiple comparisons?
  - statistical test results, e.g., P values;
  - definition of 'center values' as median or average;
  - definition of error bars as s.d. or s.e.m.

Any descriptions too long for the figure legend should be included in the methods section.

Please ensure that the answers to the following questions are reported in the manuscript itself. We encourage you to include a specific subsection in the methods section for statistics, reagents and animal models. Below, provide the page number(s) or figure legend(s) where the information can be located.

Statistics and general methods
Reported on page(s) or figure legend(s):

Corresponding Author Name: ______________________________________
Manuscript Number: ______________________________
The Public is Changing

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 Krahm, Vincent O. Tirole, Peter M. Op den Velde Boots, Andrew J. Grierson

Published: July 24, 2012  DOI: 10.1371/journal.pone.0041834
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**

**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: 19 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 items on Darwin Core. The data set can be downloaded 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.

Data published in a repository

The US Long Term Ecological Research Network


**LTER Identifier:**

knb-lter-nr.279.1

**Abstract:**

These data were collected by the Wisconsin Department of Natural Resources (WDNR) from 1967-1998. Most of these data (1967-1993) precede 1995, the year that the University of Wisconsin-NTL-LTER program took over sampling of the Yahara Lakes. However, WDNR data collected from 1997-1998 (unrelated to LTER sampling) is also included. In 1967 a joint project by the WDNR and the University of Wisconsin-Madison, Center for Limnology (CFL) was initiated on Lake Mendota. The project involved biomanipulation o...

**Owners/Creators:**

Lathrop

**Metadata:**

Select here for full metadata

**Data File(s):**

- wdnr_fyke_minifyke_seine_lengths_weights.csv
- wdnr_boomshock_lengths_weights.csv
- wdnr_pillnet_lengths_weights_93.csv
- wdnr_walleye_gasp_lengths_weights_87.csv
- wdnr_cereal_survey_lengths_weights.csv
- wdnr_cereal_survey_simpler_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.
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
Reproducible Publications and Executable Papers

\[ \text{Sweave} = \text{R} \cdot \text{LATEX} \]

\[ \text{IP[y]: Notebook} \]
Reproducible Publications

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

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

NOT published, loosely recorded:

Reproducible Publications

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

Data:
Supplementary materials, pointers to data repositories

Software:
Data preparation, data analysis, and visualization

Provenance and Workflow:
Workflow/scripts describing dataflow, codes, and parameters
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
Sharing Detailed Research Data Is Associated with Increased Citation Rate

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

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.
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

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)

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

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
Problems with Current Practice

- Data is often not made available in publications
- Limited reproducibility

Data made available through investigator’s URL
- URL does not resolve (i.e., ‘rotten’)

**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.

**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 Klein1, 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
Data: 19 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 items on Darwin Core. The data set can be downloaded 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.


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:**

lkd-lter-rd.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 Æ NTL-LTER program Æ took over sampling of the Yahara Lakes. However, WDNR data collected from 1997-1998 (unrelated to LTER sampling) is also included. In 1987 a joint project by the WDNR and the University of Wisconsin-Madison, Center for Limnology (CFL), was initiated on Lake Mendota. The project involved biomonitoring a...

**Owner/Creators:**

Lathrop

**Metadata:**

Select here for full metadata

**Data File(s):**

- wdnr_fyke minifyke seine lengths weights.csv
- wdnr_boomershock lengths weights.csv
- wdnr Gillnet lengths weights 83.csv
- wdnr_wallwe_pig lengths weights 87.csv
- wdnr_crawl_survey lengths weights.csv
- wdnr_crawl_survey simple counts.csv
Goals of this Section

1. Understand best practices
2. Understand how to implement those best practices
### Making Data Accessible: Overview of Best Practices

#### Highly connected drug file

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

#### Categories

- Computational Biology

#### Authors

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

#### Tags

- results
- tb-drugome

#### License

- CC-BY

### Description

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

### Links

Best Practices (1 of 5)

Publication in a shared repository

General minimal metadata

Domain metadata

Unique persistent identifier (PID)

Citation preference
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.arqbys.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
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,…

Credit: Kerstin Lehnert, LDEO, Columbia U.
Best Practices (2 of 5)

1. Publication in a shared repository
2. General minimal metadata
3. Domain metadata
4. Unique persistent identifier (PID)
5. Citation preference
Minimal Metadata

**General**

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

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

General

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

Recommended: CC-BY and CC0

http://creativecommons.org/licenses/
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

**Description**

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

**Links**

- http://purl.org/net/tb-drugome-run
# 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*
**SEARCHING AND BROWSING METADATA**

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

**DATA**

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

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 into an analysis environment. The WaterOneFlow Web Services format the data as the type of XML described above, WaterML 1.1.

USGS Data Source
Streamflow gauges
Network
Sites Observation Series
Values

WaterOneFlow Services
Return network information, and variable information within the network
Return site information, with a series catalog of variables measured at a site and their period of measurement
Return time series of values
{Value, Time, Qualifier}

https://www.cuahsi.org/Standards
### Best Practices (4 of 5)

#### Publication in a shared repository

A highly connected drug file published on Figshare.

#### General minimal metadata

- Categories: Computational Biology
- Authors: Daniel Garito, Lei Xie, Yinliang Zhang, Yolanda Gil, Li Xie, Sarah Kinnings, Phil Bourne
- License: CC-BY

#### Domain metadata

- File size: 4.96 KB
- Published on 20 Aug 2013 - 12:44 (GMT)

#### Unique persistent identifier (PID)

Cite this: Garito, 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.778887

#### Citation preference

Links:
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!!
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 e.g: http://www.wisc.edu/myadvisorsgroup/awesomedata.html

- Always refer to your data with the same PURL: 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

- It is easy to create your own PURLs, just remember to update whenever you move the data
DOIs can only be issued by a DOI authority (e.g., 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.0000057 or 10.1371/journal.pbio.000005.g005. Although some of you may already be acquainted with DOIs, others of you may wonder what they are, how they are used, and why we are using them.

**What Are DOIs?**

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)

1. Publication in a shared repository
2. General minimal metadata
3. Domain metadata
4. Unique persistent identifier (PID)
5. Citation preference
Data Citation Format

Data repositories and journals often specify how to cite data.
Versatility of Data Repositories

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.
**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 “microattributions” 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

https://www.flickr.com/photos/vizzual-dot-com/2655969483/
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 requirements, another researcher can reproduce the experiments described, (2) our aim to promote openness. PLOS journals can be built upon by future researchers. Therefore, if new software or a new version 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](http://www.opensource.org/licenses)).

- **Documentation for running and installing the software.** For end-user applications, this is a prerequisite; for software libraries, instructions for using the application program interface are preferred.

- **A test dataset with associated control parameter settings.** Where feasible, results on 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 for creating user accounts, logging in or otherwise registering personal details. The repository should include 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 source code.
Some journal articles describe a piece of software

Some publications have “software papers” or “software metapapers”
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!”
“Scientists and engineers spend more than 60% of their time just preparing the data for model input or data-model comparison” (NASA A40)

"Dark Software" is the counterpart of "Dark Data" [Heidorn 2008]

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

https://www.flickr.com/photos/vizzual-dot-com/2655969483/
Some Notes on Making Software Ready for Publication

- Source code vs executable
- Making software run elsewhere
- Making software modular
- Making software configurable
- Making software report errors
- Providing test data
- 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
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
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
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
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
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

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

Version Control

Community Contributions
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

Garijo, Daniel; Xie, Lei; Zhang, Yinliang; Gil, Yolanda; Xie, Li (2013) Tool for computing anomalies, GitHub. V.1
http://dx.doi.org/10.5281/zenodo.18765
Retrieved 11:05, Feb, 15, 2015 (GMT)
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

```javascript
function enEdition()
{
    /* Ne rien faire mode edit */
    if ( encodeURIComponent(document.title) !== '')
        return; // /&preload=/

    if (!wgPageName.match(/Discussion|Discussion/))
    {
        var diff = new Array();
        var status = document.getElementById('diff-status');
        var p = document.getElementById('p-diff');
        var avancementTraduction; var avance;

        /* ++++++++++++ Parser ++++++++++++ */
        var params = document.location.search.substring(1).split('&');
        var i = 0;
        var tmp; var name;
        while ( i < params.length )
        {
            tmp = params[i].split('=');
            name = tmp[0];
            switch( name ) {
                case 'status':
                    status = {'ma': 0, 'st': 1, 'cm': 2, 'vb': 3, 'c': 4, 'r': 5, 'e': 6, 'x': 7};
                break;
            }
            var a = document.createElement('a');
            a.appendChild(doc.createTextNode(name + ' ' + status[name]));
            var li = document.createElement('li');
            li.appendChild(a);
            diff.push(li);
        }

        status = document.getElementById('status');
        status.innerHTML = '...
```
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

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

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
Software entries from distributed repositories are readily accessible.

Semantic search

Comparison matrix of software entries

Metadata completion highlighted

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…”

*Leci n’est pas une pipe.*
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

<table>
<thead>
<tr>
<th>Description</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short description of this repository</td>
<td>Website for this repository (optional)</td>
</tr>
</tbody>
</table>

- **5 commits**
- **1 branch**
- **1 release**
- **1 contributor**

**Update README.md**

- **LICENSE**
- **README.md**
- **dudt.ncl**
- **dudt_runave.ncl**
- **fv.ncl**
- **grib2netcdf.csh**
- **pgf.ncl**
- **plot_pgf_x.ncl**
- **plot_ududx_runave.ncl**

**Commit Details**

- Initial commit
- Update README.md
- Add all ncl
- Grib2nc
- Add all ncl
- Add all ncl
- Add all ncl
- Add all ncl
- Add all ncl

**Last Commit**

- a35bf619e5

**GitHub Repository**

- **HTTPS clone URL**
  - https://github.com/jihyun
- **Clone In Desktop**
- **Download ZIP**
Describing Software with OntoSoft

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

Automatic crawlers import metadata from code repositories (eg GitHub)
Finding Software with OntoSoft

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

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

<table>
<thead>
<tr>
<th>Software List</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Name</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DrEICH algorithm</td>
<td></td>
<td>EDIT</td>
</tr>
<tr>
<td>PIHM</td>
<td></td>
<td>EDIT</td>
</tr>
<tr>
<td>PIHMgis</td>
<td></td>
<td>EDIT</td>
</tr>
<tr>
<td>TauDEM</td>
<td></td>
<td>EDIT</td>
</tr>
<tr>
<td>WBMsed</td>
<td></td>
<td>EDIT</td>
</tr>
</tbody>
</table>

Filter Software List

Search

Author

Keywords: Hydrological model OR Hydrology

Language: C++

License: GNU General Public License v2.0

GNU General Public License
Comparing Alternatives with OntoSoft

## Compare Software

<table>
<thead>
<tr>
<th>Software</th>
<th>PIHM</th>
<th>PIHMgis</th>
<th>DrEICH</th>
<th>TauDEM</th>
<th>WBMsed</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Compare Software</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>DrEICH algorithm, PIHM, PIHMgis, TauDEM, WBMsed</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### What domain specific keywords for this software?
- (eg: hydrology, climate)
- Geomorphology, Hydrological, Bedrock channel erosion
- Basins, Continental
- Basins, GIS
- Hydrologically corrected DEM, Watershed
- Sediment flux, Global model, Hydrological model

### What Operating Systems can the software run on?

<table>
<thead>
<tr>
<th>Software</th>
<th>Operating Systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>PIHM</td>
<td>Unix, Linux</td>
</tr>
<tr>
<td>PIHMgis</td>
<td>Unix, Windows, Linux, Mac OS</td>
</tr>
<tr>
<td>DrEICH</td>
<td>Unix, Windows, Linux, Mac OS</td>
</tr>
<tr>
<td>TauDEM</td>
<td>Unix, Windows, Linux, Mac OS</td>
</tr>
<tr>
<td>WBMsed</td>
<td>Unix, Linux</td>
</tr>
</tbody>
</table>

### Is there any test data available for the software?

<table>
<thead>
<tr>
<th>Software</th>
<th>Test Data Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>PIHMgis</td>
<td><a href="http://sourceforge.net/projects/pihmmodel/">http://sourceforge.net/projects/pihmmodel/</a></td>
</tr>
<tr>
<td>DrEICH</td>
<td></td>
</tr>
<tr>
<td>TauDEM</td>
<td><a href="http://csdms.colorado.edu/wiki/Model:TauDEM#Testing">http://csdms.colorado.edu/wiki/Model:TauDEM#Testing</a></td>
</tr>
<tr>
<td>WBMsed</td>
<td><a href="http://csdms.colorado.edu/wiki/Model:WBMsed#Testing">http://csdms.colorado.edu/wiki/Model:WBMsed#Testing</a></td>
</tr>
</tbody>
</table>

**Test Data Description:**
- Two test DEMs are included in the repository.
- Test Data Description: Upper Juniata River 875 km²: see http://sourceforge.net/projects/pihmmodel/
Publishing Software Metadata with OntoSoft

http://www.ontosoft.org/portal

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

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)

CC-BY Attribution
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]
“**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

Reaeration

Churchill

R code
What the Paper Says Versus What the Actual Software Does (from [Garijo et al 2013])

Comparison of ligand binding sites using SMAP
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

Sweave = R • LaTeX

CDF Computable Document Format
Documents come alive with the power of computation

http://ipython.org/notebook.html
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.)

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 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 Herculanum in 79 AD. Both eruptions claimed thousands of lives.

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

References

1. @ This article incorporates public domain material from https://en.wikipedia.org/wiki/Stratovolcano Retrieved 2009-01-19.
5. a b c d e f g h j k l m This article incorporates public domain material from Wikipedia. 
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."

<table>
<thead>
<tr>
<th>Title</th>
<th>Prince Larry did not take drugs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creator</td>
<td>REUTERS journalist</td>
</tr>
<tr>
<td>Subject</td>
<td></td>
</tr>
<tr>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>Publisher</td>
<td>FA Times</td>
</tr>
<tr>
<td>Contributor</td>
<td>Duckingham Palace</td>
</tr>
</tbody>
</table>

- **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

http://www.w3.org/TR/prov-primer/
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
Understanding kinematic data from the Moine thrust zone (doi:10.1016/j.ess.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) using a petrographic microscope. We rotated to a common reference frame (doi:10.6084/m9.figshare.798887) using Duyster’s StereoNett program (doi:10.5281/zenodo.18954). We plotted the data on lower hemisphere, equal area projections (doi:10.6084/m9.figshare.798887) using Duyster’s StereoNett program (doi: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**
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 (e.g., 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
Establishing Your ORCID Profile & Researcher Impact

Gail Clement
Head of Research Services, Caltech Library

http://orcid.org/0000-0001-5494-4806
gperetsm@caltech.edu

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

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/
3 Open Standards form the building blocks of today’s web Research Information Systems (RIS)

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.
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/

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

Code ‘10’ assigned to all DOIs by International DOI Federation.

Prefix assigned to publisher by DOI Registration Agency.

Suffix assigned to the work by its Publisher.

https://doi.org/10.1353/pla.2011.0031

10 DOI Registration Agencies worldwide; CrossRef used for scholarly pubs.
CrossRef 101

- 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

https://www.crossref.org/
Step 1 to do

Step 1 to do

$ cat clement2011.bib

Step 2. Overview

Represent your identity as a unique number using ORCiD so it is machine readable and actionable
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
ORCiD 101
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
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
Steps 3 + 4 Overview

Connect your citations with DOIs to your ORCiD works
Step 3 to do

Connect your citations with DOIs to your ORCiD works, manually
Step 4 to do

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

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
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

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 Cotlarciuc, Toby Andrew, Tracy Dew, Gail Clement, Raj Gill, Gabriela Surdulescu, Roy Sherwood, Kourosh R. Ahmadi
Step 4 to do
Manage duplicate entries in your ORCiD works
Step 5 Overview

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

http://openvivo.org/vis/capabilitymap#Artificial%20intelligence%7C30%7C1

VIVO is member-supported, open source software and an ontology for representing scholarship. | vivoweb.org
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
VIVO applies the Contributor Role taxonomy developed through Project CRediT
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/


“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!
The Scientific Paper of the Future: An Author Checklist

OntoSoft Training

Part 7

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

http://www.sciencicpaperofthefuture.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.
**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
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

- **Authors:**
- **Date of publication:**
- **Time of retrieval:**
- **Permanent unique identifier:**
- **Name:**
- **Repository:**
Datasets should have general-purpose metadata specified (creator, date, name, etc.)

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

- 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

```javascript
function edition {
  /* Ne rien faire mode edit */
  if( encodeURIComponent(document)
      turn;
      // /&preload=/

  if ( !wgPageName.match(/Discussion
      var diff = new Array();
      var status; var pecTraduction; var p
      var avancementTraduction; var advance

  /* ******************** Parser ******************** */
  var params = document.location.search
      .split('&');
  var i = 0;
  var tmp; var name;
  while ( i < params.length )
  {
    tmp = params[i].split('=');
    name = tmp[0];
    switch( name ) {
      case 'status':
        status = tmp[1];
    }

https://commons.wikimedia.org/wiki/File:Source_code_in_Javascript.png
```
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

function enEdition()
{ /* Ne rien faire mode edit */
  if( encodeURIComponent(document.title)

  return;
  // /&preload=/

  if ( !wgPageName.match(/Discussion|Discussion/) )
    var diff = new Array();
    var status; var pecTraduction; var p;
    var avancementTraduction; var avance;

  /* *********** Parser *********** */
  var params = document.location.search

  i = 0;
  var tmp; var name;
  while ( i < params.length )
  {
    tmp = params[i].split('=');
    name = tmp[0];
    switch( name )
    {
      case 'status':
        status = tmp[1];
      

https://commons.wikimedia.org/wiki/File:Source_code_in_Javascript.png
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

Garijo, Daniel; Xie, Lei; Zhang, Yinliang; Gil, Yolanda; Xie, Li (2013) Tool for computing anomalies, GitHub. V.1
http://dx.doi.org/10.5281/zenodo.18765
Retrieved 11:05, Feb, 15, 2015 (GMT)
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 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 tightly coupled to a GIS interface. PIHMgis which is open source, platform independent and extensible. The tight coupling between GIS and the model is achieved by developing a shared data-model and hydrologic-model data structure.

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

- What are general categories (keywords, labels) for this software?
  - Hydrology
  - Basins
  - Continental

- Is there a project website for the software?
  - http://www.pihm.psu.edu/pihm_home.html

**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
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
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 (e.g., figshare, myExperiment, etc)

6. Get a unique persistent identifier for the workflow, the provenance, or both
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
Understanding kinematic data from the Moine thrust zone (doi:10.1016/j.ess.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) using a petrographic microscope. We rotated to a common reference frame (doi:10.6084/m9.figshare.798887) using Duyster’s Stereonett program (doi:10.5281/zenodo.18954). We plotted the data on lower hemisphere, equal area projections (doi:10.6084/m9.figshare.798887) using Duyster’s Stereonett program (doi: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
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

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
Rainfall
Snow melt
Snow fall
Recharge
Groundwater discharge
Runoff
Transpiration
Canopy Evap.
Soil Evap.
Unsaturated Zone
Saturated Zone