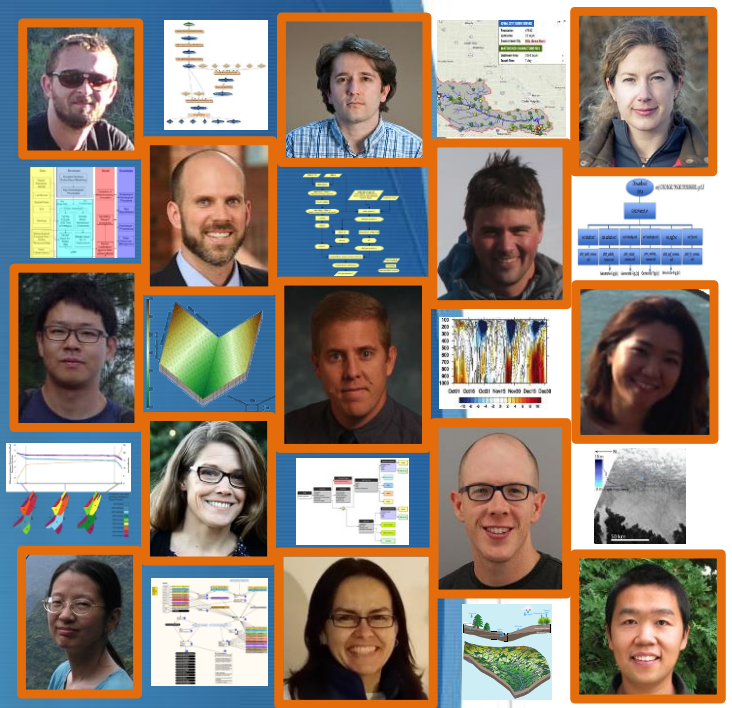
The Scientific Paper of the Future



http://www.scientificpaperofthefuture.org

Jan, 2023. Universidad Politécnica de Madrid

Instructor: **Daniel Garijo** with slides from **Yolanda Gil**

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

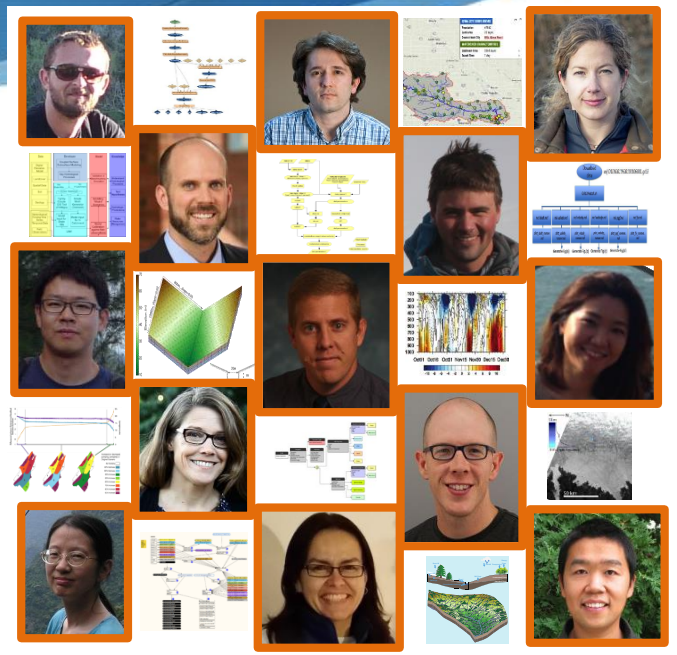
ICER-1440323
ICER-1343800

Geoscience Paper of the Future (GPF) Initiative

ICER-1440323
ICER-1343800

http://www.ontosoft.org/gpf

- Motivation: Scientists want to learn best practices for software sharing, but prefer to do it while doing research
  - Train paper authors on best practices as they write a Geoscience Paper of the Future (GPF)
- Proposed by members of the OntoSoft Early Career Advisory Committee (~30 members)
  - Covering diverse areas of geosciences



- Training: Developed a 3 hour training session
- Journal Special Issue:
  - Write a GPF about new research being done
  - Write a GPF to document an already published paper

AGU American Geophysical Union
Earth and Space Science
AN OPEN ACCESS AGU JOURNAL
Special Section: Geoscience Papers of the Future

www.scientificpaperofthefuture.org

The Scientific Paper of the Future Initiative

Home | Motivation | What is a SPF | Sessions | Materials | Events | Gallery | FAQ | Organization

"Towards the Geoscience Paper of the Future: Best Practices for Documenting and Sharing Research from Data to Software to Provenance" Gil et al, Earth and Space Science, 2016.
http://dx.doi.org/10.1002/2015EA000136

SEG SOCIETY OF EXPLORATION GEOPHYSICISTS

Geophysics: Special Issue on Geoscience Papers of the Future

AGU American Geophysical Union

Earth and Space Science
AN OPEN ACCESS AGU JOURNAL

Special Section: Geoscience Papers of the Future

AI Magazine, Vol. 39, No. 3, Fall 2018. http://doi.org/10.1609/aimag.v39i3.2816

On Reproducible AI: Towards Reproducible Research, Open Science, and Digital Scholarship in AI Publications

Odd Erik Gundersen, Norwegian University of Science and Technology
Yolanda Gil, University of Southern California
David W. Aha, US Naval Research Laboratory

NATURE REVIEWS | NEUROSCIENCE

Scanning the horizon: towards transparent and reproducible neuroimaging research

Russell A. Poldrack[1], Chris I. Baker[2], Joke Durnez[1,3], Krzysztof J. Gorgolewski[1], Paul M. Matthews[4], Marcus R. Munafò[5,6], Thomas E. Nichols[7], Jean-Baptiste Poline[8], Edward Vul[9] and Tal Yarkoni[10]

Towards the neuroimaging paper of the future

In this Analysis article, we have outlined a number of problems with current practice and made suggestions for improvements. Here, we outline what we would like to see in the neuroimaging paper of the future, inspired by related work in the geosciences[71].

Why Learn to Write a Scientific Paper of the Future

1. Practice **open science and reproducible research**
2. **Get credit** for all your research products
   - Citations for software, data, containers, notebooks, samples, …
3. **Increase citations** of your papers
4. Write impressive **Data Management Plans**
5. **Extend your CV** with data and software sections
6. Improve the **management of your research assets**
7. **Reproduce** your work from years ago and build on it
8. Address new **funder and journal requirements**
9. Attract **transformative students**
10. Demonstrate **leadership** by stepping into the future

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
- **Recommendations that are mindful of effort required**
  - 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

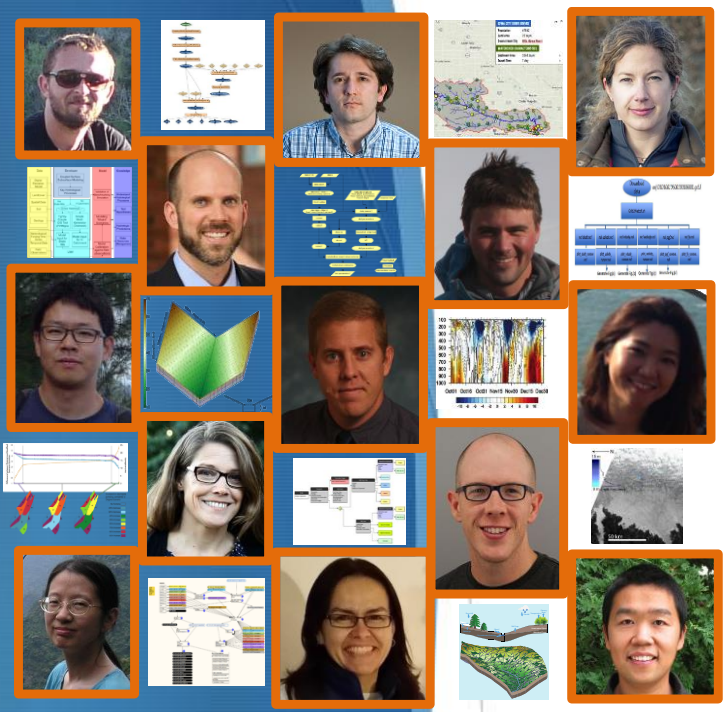
Part I

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

Part II

1. Documenting software dependencies
2. Documenting methods and workflows
3. Documenting provenance
4. *PRACTICAL EXERCISE*
5. Summary of author checklist

The Scientific Paper of the Future: Motivation and Overview



Part 1.1

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

http://www.scientificpaperofthefuture.org

ICER-1440323
ICER-1343800

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

figshare

DRYAD

The Dataverse Network™ Project

Open access

creative commons

open source initiative

Open publications

PLOS

Open source

github
SOCIAL CODE HOSTING

Apache

sourceforge

wikigenes
collaborative publishing

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 (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

http://www.nature.com/nature/newspdf/metrics_survey.pdf

The Science Community is Changing

zenodo

figshare

The Dataverse Project

DRYAD

PANGAEA

W3C® PROV

RDA RESEARCH DATA ALLIANCE

SIP

PLOS

Impactstory

depsy

Universities are Changing: Major Initiatives in Data Science

MIT HAYSTACK OBSERVATORY

Carnegie Mellon University ML MACHINE LEARNING DEPARTMENT

Data Science Institute COLUMBIA UNIVERSITY

NORTHEAST BIG DATA INNOVATION HUB

WEST BIG DATA INNOVATION HUB

JOHNS HOPKINS INSTITUTE FOR DATA-INTENSIVE ENGINEERING & SCIENCE

MIDAS MICHIGAN INSTITUTE FOR DATA SCIENCE UNIVERSITY OF MICHIGAN

CENTER FOR DATA-DRIVEN DISCOVERY Caltech

DSRC DATA SCIENCE RESEARCH CENTER Rensselaer

The New York Times

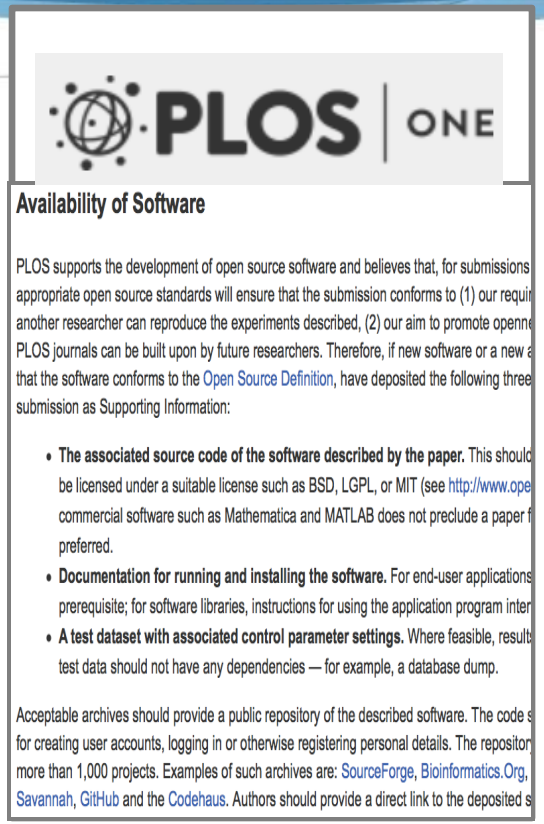
Program Seeks to Nurture 'Data Science Culture' at Universities

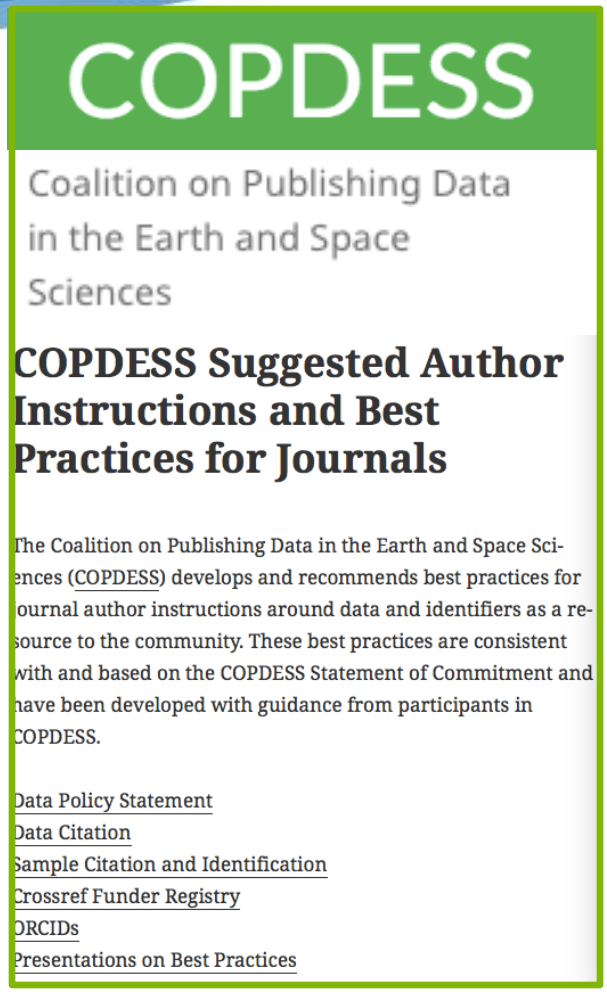
By STEVE LOHR
NOVEMBER 12, 2013

UNIVERSITY of WASHINGTON Berkeley UNIVERSITY OF CALIFORNIA NYU

three universities and supported by $37.8 million in funding from the Moore Foundation and the Sloan Foundation. The three universities in the partnership are New York University, the University of Washington and the University of California, Berkeley. The program is being announced today in Washington at an event organized by the White House Office of Science and Technology Policy, to

Publishers Are Changing Guidelines for Authors





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;

The Public is Changing: Interest in Doing Science

ZOONIVERSE
REAL SCIENCE ONLINE

foldit
Solve Puzzles for Science

oldWeather

eBird

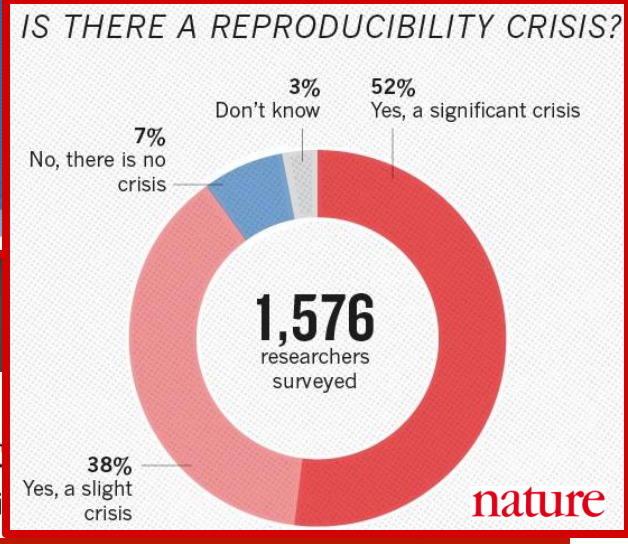
PLOS | ONE

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

Reproducibility



Illuminating the black box

nature

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

Friday, December 2, 2011 As of 12:00 AM New York 43° | 34°

THE WALL STREET JOURNAL. HEALTH

HEALTH INDUSTRY | DECEMBER 2, 2011

Scientists' Elusive Goal: Reproduc

n September, Bayer published a study describing how i

COMPUTER SCIENCE

Accessible Reproduc

Science

putation in research grows, needed to expand recording,

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

statistics journals

r sensational

s. The most recent issue of the Annals of Applied Statistics is an

lleges that cancer patients

Human lives

The New York Times

NYTimes: Home - Site Index - Archive - Help

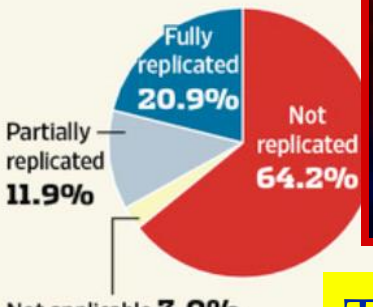
Science

Reliability

Nobel Laureate Retracts Two Paper

No Cure

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



Source: Nature Reviews Drug Discover

Financial

The New York Times

Retracted Scientific Studies: A Growing List

Scientific integrity

Trust

Reproducible Publications and Executable Papers

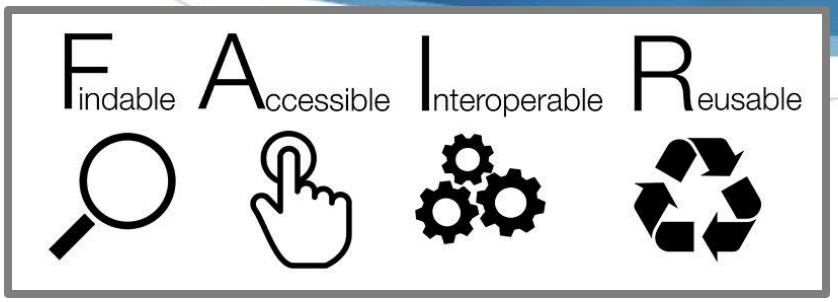
$$\text{Sweave} = \text{R} \cdot \text{LaTeX}$$

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

The FAIR Principles

https://www.force11.org/group/fairgroup/fairprinciples

doi.org/10.1038/sdata.2016.18



To be Findable:

F1. (meta)data are assigned a globally unique and eternally persistent identifier.
F2. data are described with rich metadata.
F3. (meta)data are registered or indexed in a searchable resource.
F4. metadata specify the data identifier.

TO BE ACCESSIBLE:

A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
A1.1 the protocol is open, free, and universally implementable.
A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
A2 metadata are accessible, even when the data are no longer available.

TO BE INTEROPERABLE:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles.
I3. (meta)data include qualified references to other (meta)data.

TO BE RE-USABLE:

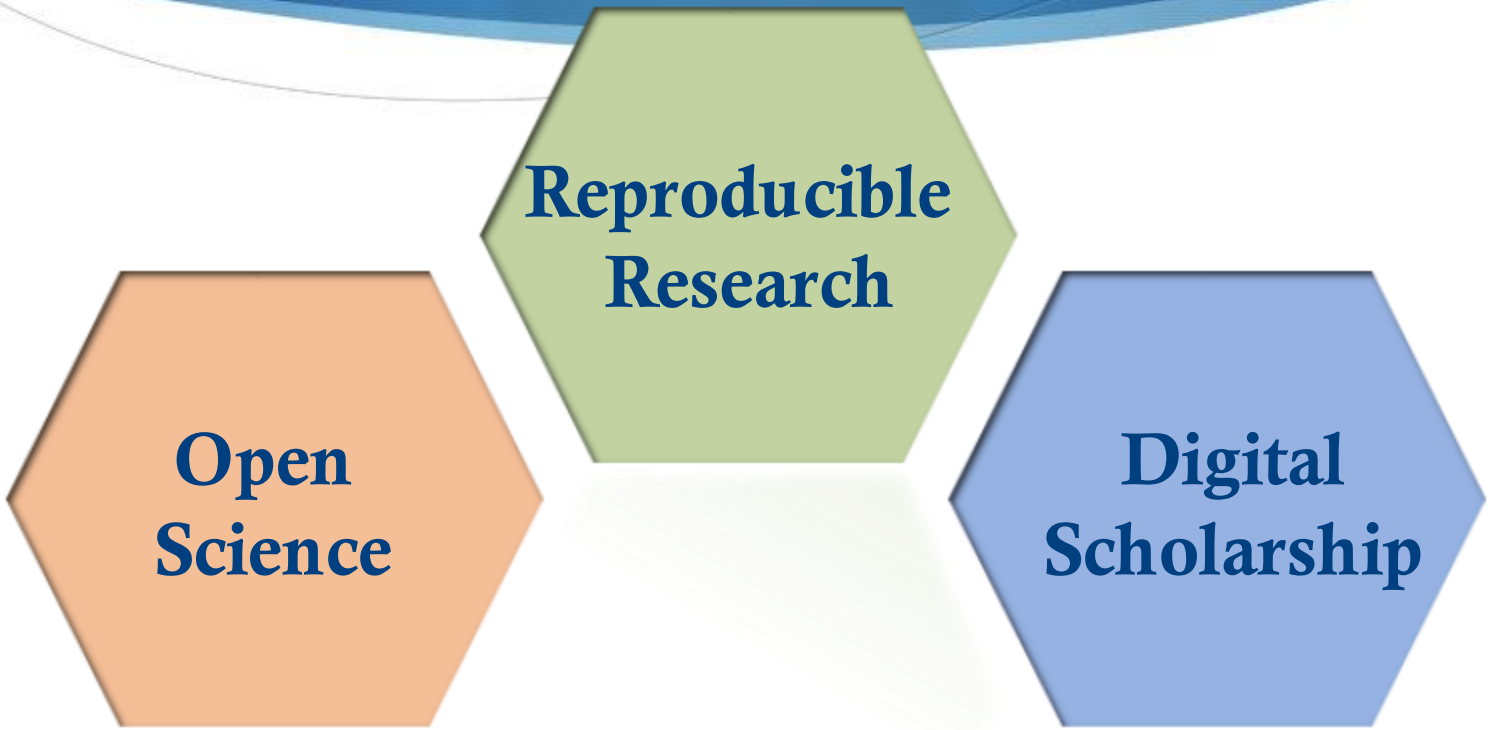
R1. meta(data) have a plurality of accurate and relevant attributes.
R1.1. (meta)data are released with a clear and accessible data usage license.
R1.2. (meta)data are associated with their provenance.
R1.3. (meta)data meet domain-relevant community standards.

G20 Leaders' Communique Hangzhou Summit

Hangzhou, 5 September 2016

"We support appropriate efforts to promote open science and facilitate appropriate access to publicly funded research results on findable, accessible, interoperable and reusable (FAIR) principles."

Core Recommendations for Scientific Publications



CODATA

RDA RESEARCH DATA ALLIANCE

DataCite

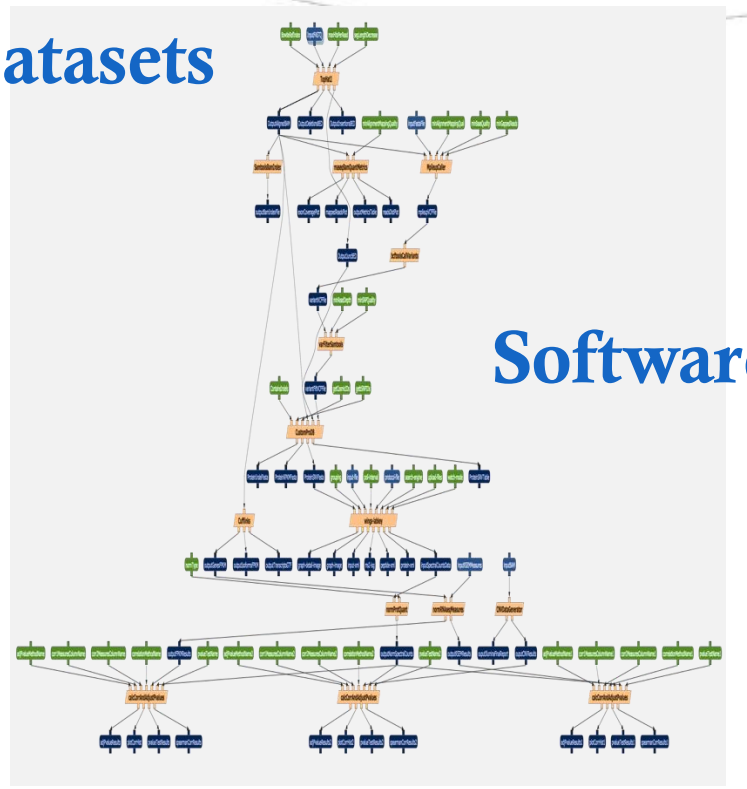
FORCE11 The Future of Research Communications and e-Scholarship

SIP

1) Reproducible Research

Datasets

Software



Workflow

Hypotheses

Name

Protein kinase C delta-binding protein is expressed in patient sample

Lines of Inquiry

Name

Protein association with patient

Description

This line of inquiry is used for protein->patient association

Query (Ctrl-Space for suggestions)

```
?x :expressedIn ?sample

?sample :collectedFrom ?p
?p a :Patient
?sample a :Sample
?e1 :experimentedOn ?sample
?e2 :experimentedOn ?sample
FILTER(?e1 != ?e2)
?e1 :produceData ?d1
?e2 :producedData ?d2
?d1 a :MassSpecData
?d2 a :RNASeq
```

Workflows to Run +

proteogenomic_analysisBasic

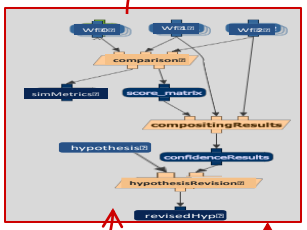
Variable Bindings: InputFASTQ = ?d2, input-file = ?d1

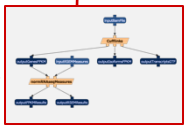
proteomics_analysis

Variable Bindings: input-file = ?d1

Meta-Workflows to Run +

Protein_Diff_WF





Experimental Design

2) Open Science

Shared repositories

zenodo

The Dataverse Project

figshare

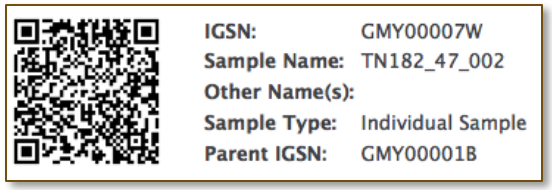
github SOCIAL CODE HOSTING

CIG COMPUTATIONAL INFRASTRUCTURE for GEODYNAMICS

Persistent unique identifiers

PURL

DOI



Licenses

creative commons

CC BY

open source initiative

3) Digital Scholarship

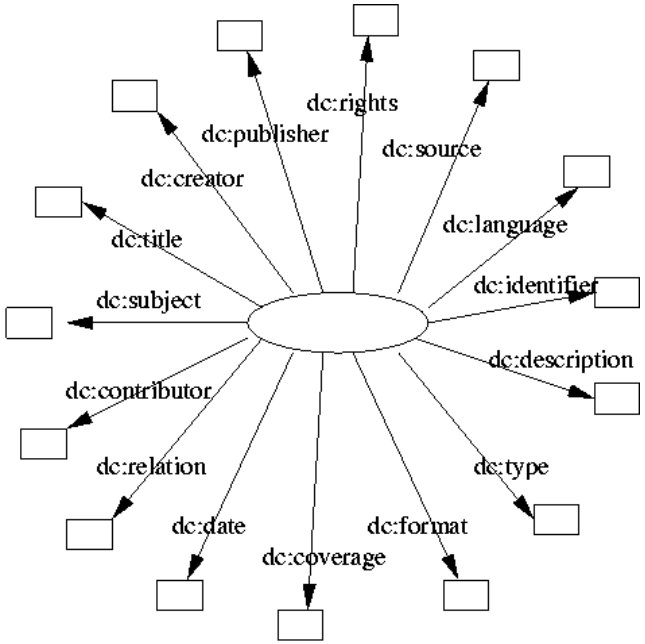
Data and software citation

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)

Persistent unique identifier

Repository

Metadata



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 Research

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)

Reproducible Articles

Modern Published Articles

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

Data:
Supplementary materials,
pointers to data repositories

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

Beyond Reproducible Publications

Reproducible Publications

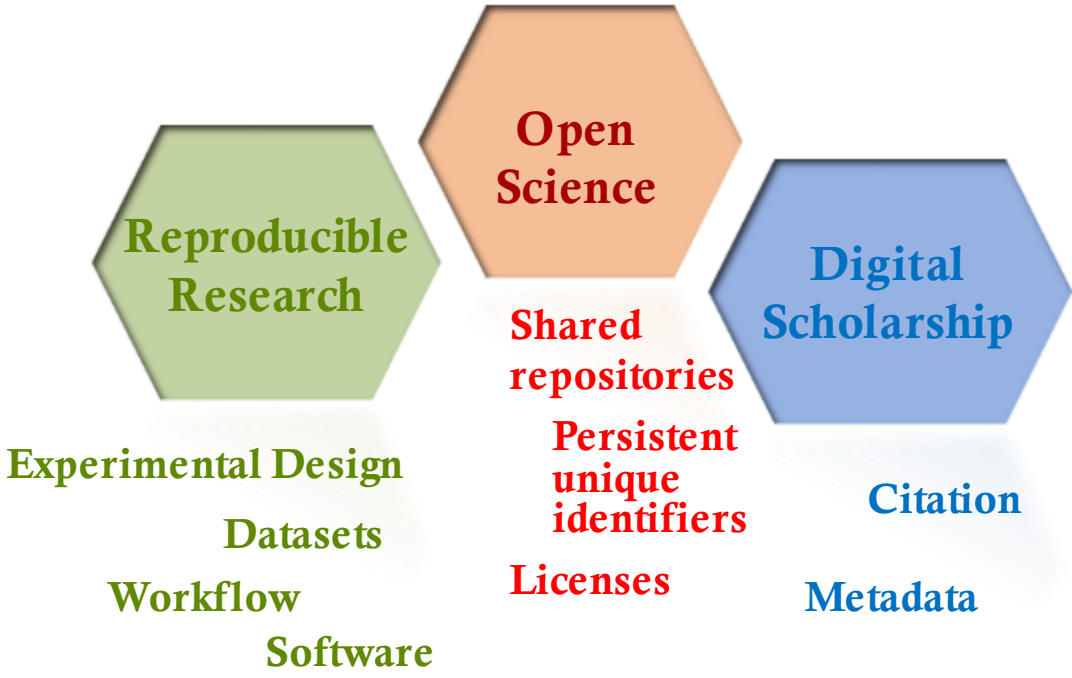
Text:
Narrative of method,
the data is in tables, figures/plots,
the software used is mentioned

Data:
Supplementary materials,
pointers to data repositories

Software:
Data preparation,
data analysis, and visualization

Provenance and methods:
Workflow/scripts describing
dataflow, codes, and parameters

The Scientific Paper of the Future has further requirements



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 Training

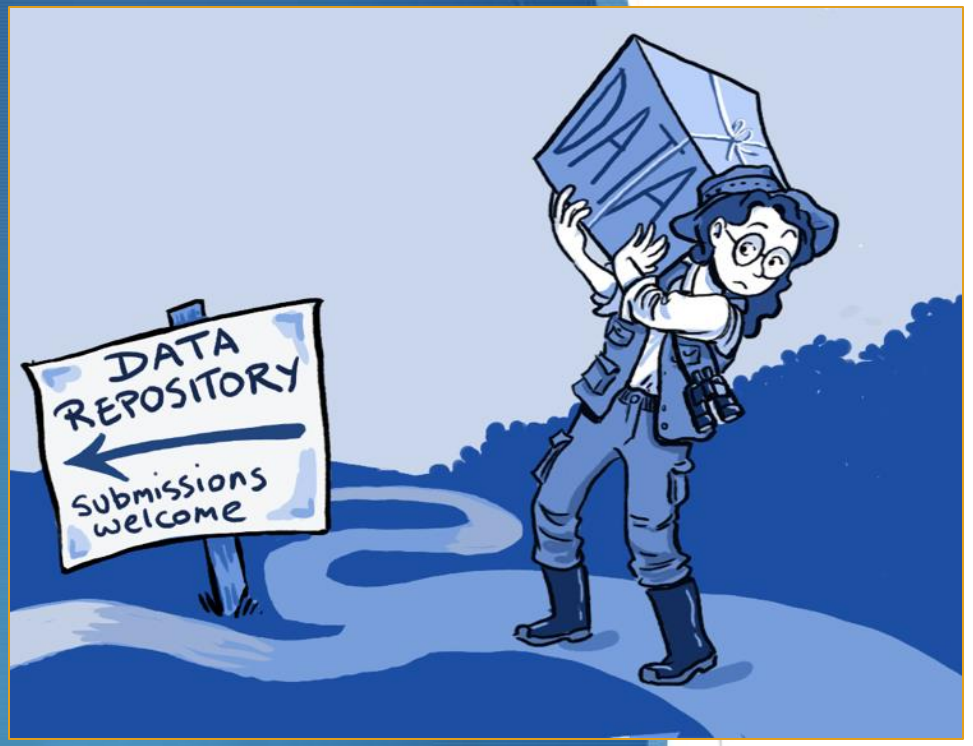
Part I

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

Part II

1. Documenting software dependencies
2. Documenting methods and workflows
3. Documenting provenance
4. *PRACTICAL EXERCISE*
5. Summary of author checklist

Data in the Scientific Paper of the Future



Part 1.2

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

SCIENTIFIC Onto Soft PAPER OF THE FUTURE

Onto Soft

NSF ICER-1440323
ICER-1343800

http://www.scientificpaperofthefuture.org

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 Sompel[1], Robert Sanderson[1], Harihar Shankar[1], Lyudmila Balakireva[1], Ke Zhou[2], Richard Tobin[2]

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

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

Owners/Creators:

Lathrop

Metadata:

Select here for full metadata

Data File(s):

- wdnr_fyke_minifyke_seine_lengths_weights.csv
- wdnr_boomshock_lengths_weights.csv
- wdnr_gillnet_lengths_weights_93.csv
- wdnr_walleye_age_lengths_weights_87.csv
- wdnr_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	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,bp
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,ptl
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,Rv
Daunorubicin	44	12	mmaA4,Rv1264,thyX,lppX,secA1,serA1,Rv3529c
Triclosan	42	5	pepD,Rv1264,thyX,ethR,trxB2,
Darunavir	40	15	pknD,pepD,fabH,Rv1941,devB,ppp,ftsZ,cyp125

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

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

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

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,trxB2,
Darunavir	40	15	pknD,pepD,fabH,Rv1941,devB,ppp,ftsZ,cyp12

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

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 Bourn

Tags

- results
- tb-drugome

License (what's this?)

CC-BY

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

“Dark Data”

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

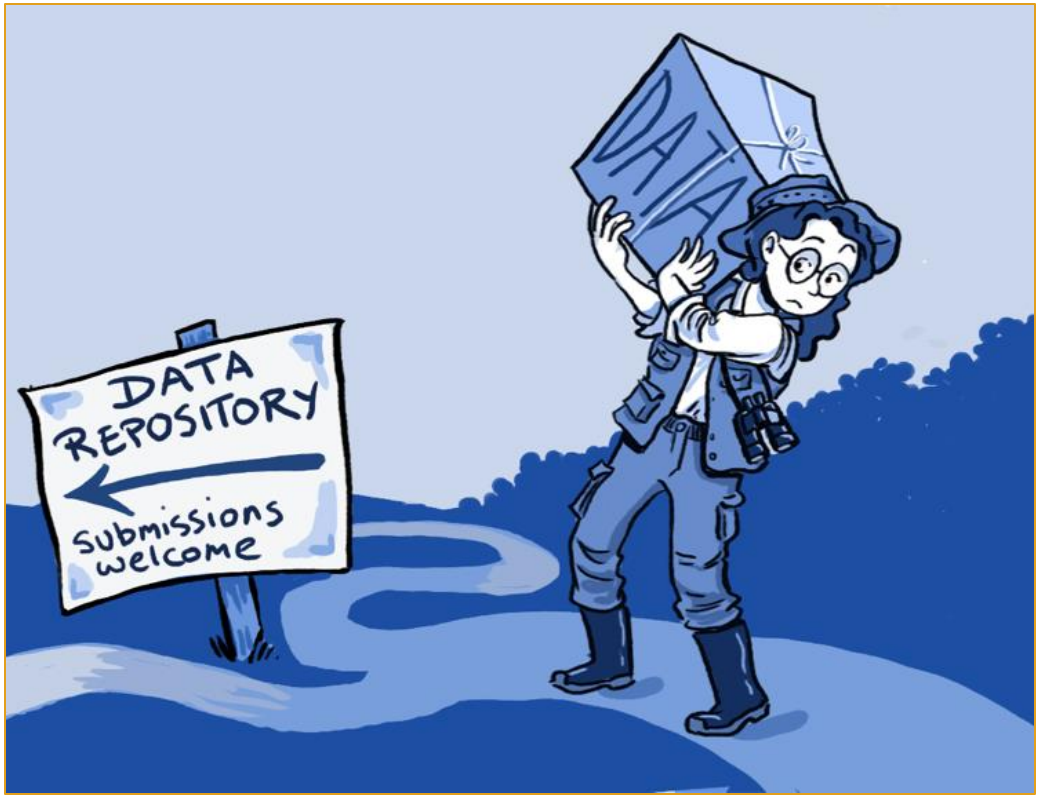
P. Bryan Heidorn

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.

Discoverability through Shared Repositories and Metadata for Data and Software



Popular Data Repositories

Not Curated

Curated

zenodo

figshare

DRYAD

The Dataverse Project

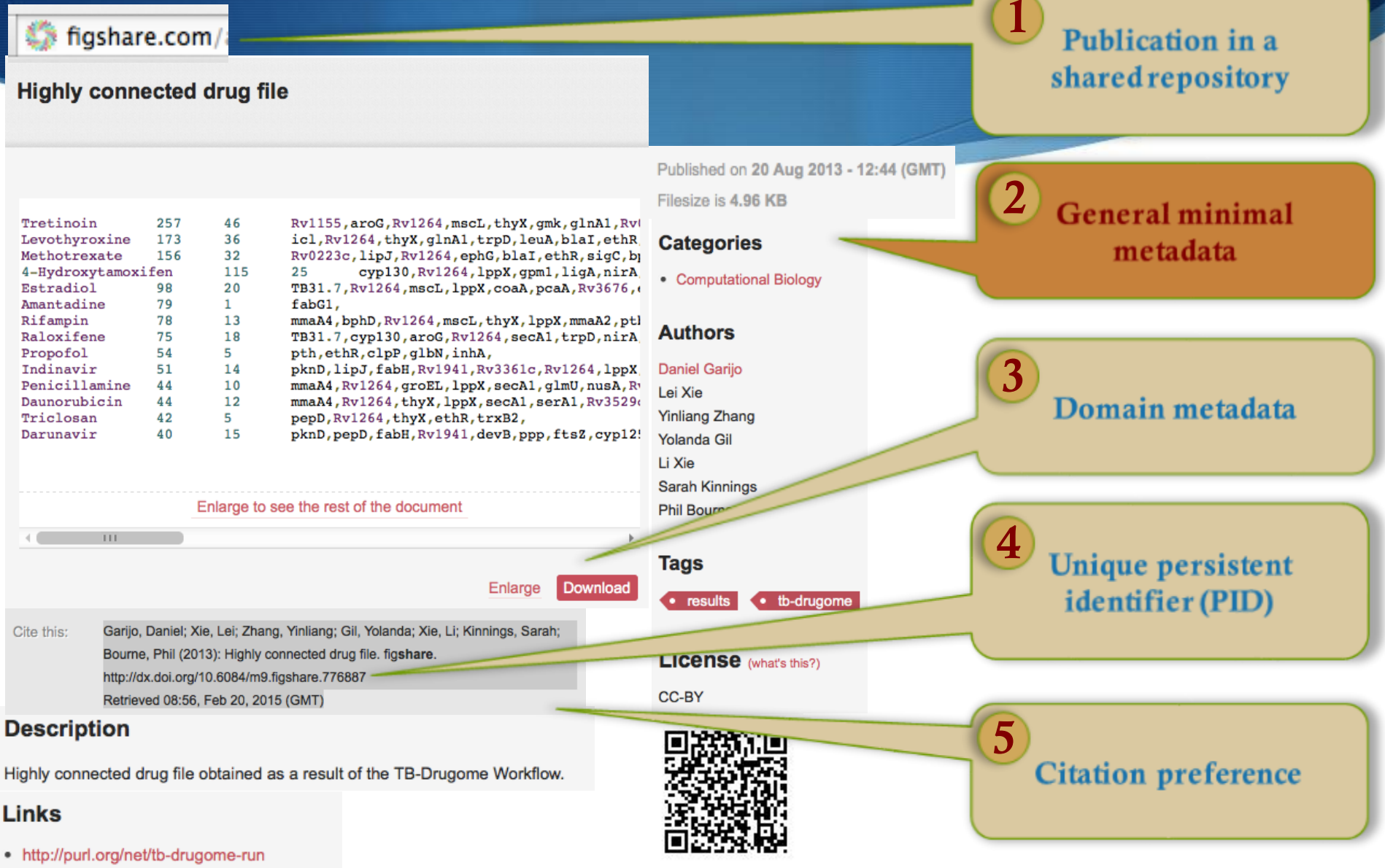
PANGAEA

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

http://www.thestaffingstream.com/2012/08/06/the-buzz-about-talent-communities/

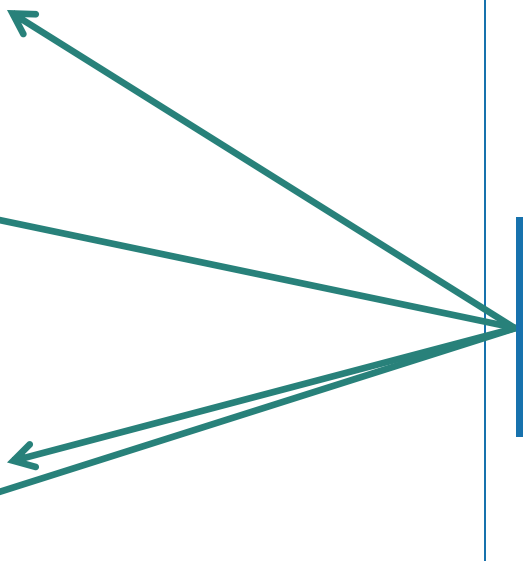
Best Practices (2 of 5)



Minimal Metadata

General

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



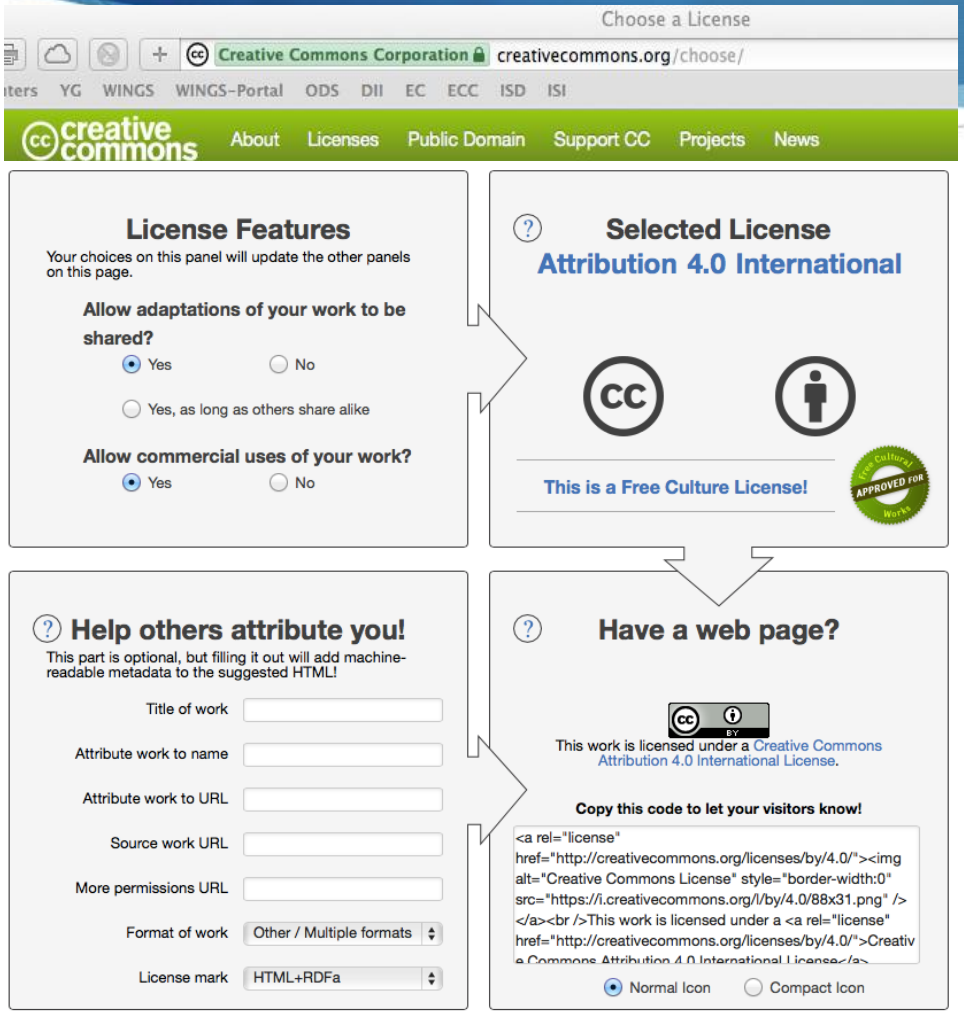
Typical of digital libraries,
e.g. the Dublin Core standard
(http://dublincore.org/documents/dcmi-terms/)

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

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.

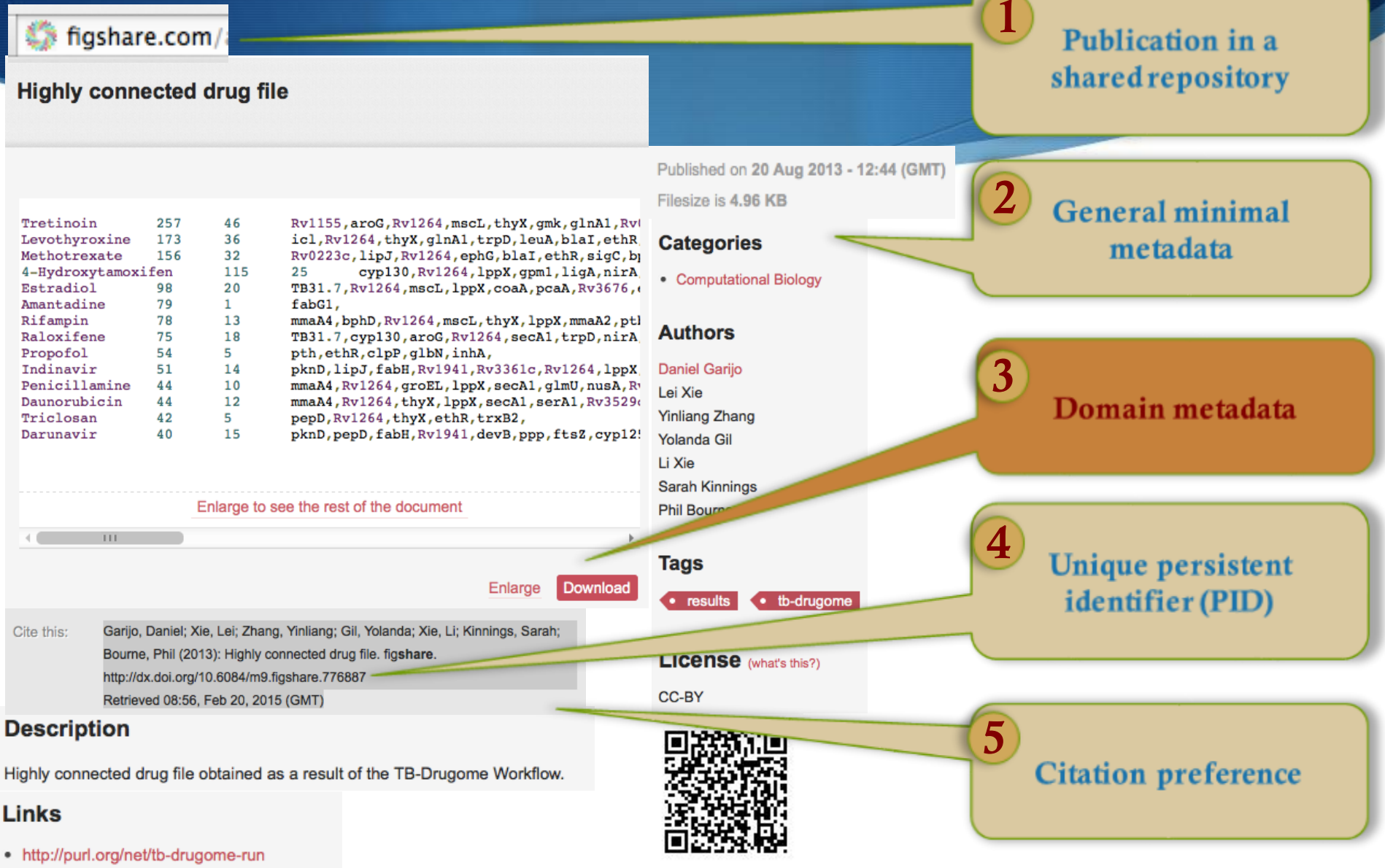
CC0 (datasets) "No rights reserved"

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/

Best Practices (3 of 5)



Domain-Specific Metadata

General

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

Domain Specific

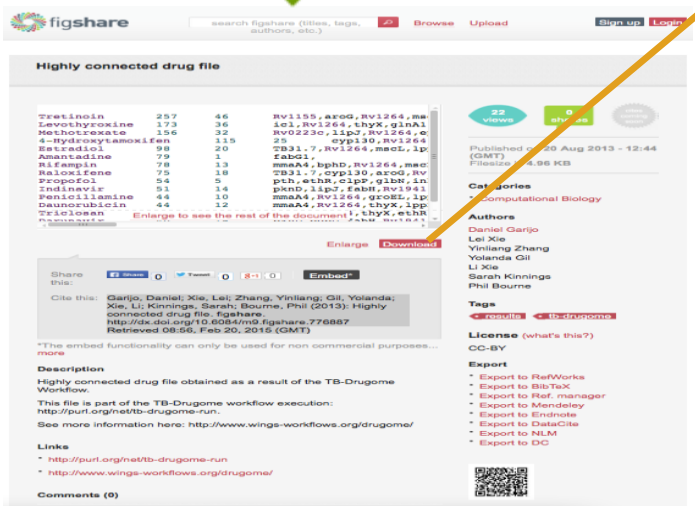
- Collection information
- Pre-processing
- Dataset characteristics

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_dr[illegible]_file/776887

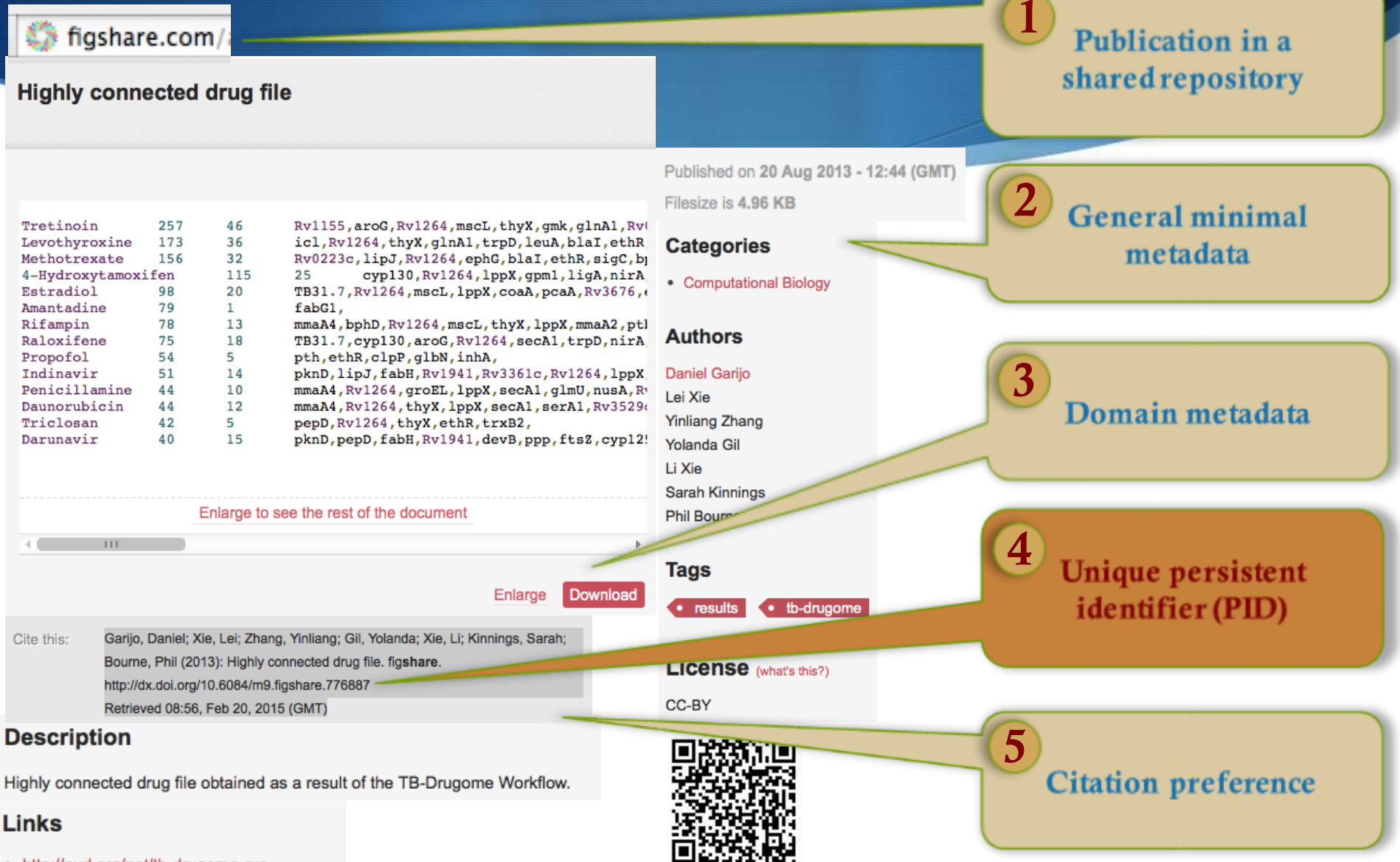


DATA

- http://files.figshare.com/1175525/highlConnectedDrugs.txt

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,trxB2,
Darunavir	40	15	pknD,pepD,fabH,Rv1941,devB,ppp,ftsZ,cyp12

Best Practices (4 of 5)



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, or other PURL services
    - Create a PURL, and direct it to where you 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

Digital Object Identifier (DOI)

PLoS Biol. 2003 Nov; 1(2): e57.
Published online 2003 Nov 17. doi: 10.1371/journal.pbio.0000057

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

The What and Whys of DOIs

Susanne DeRisi, Rebecca Kennison, and Nick Twyman

Data repositories can issue DOIs for data

DOIs are **free**

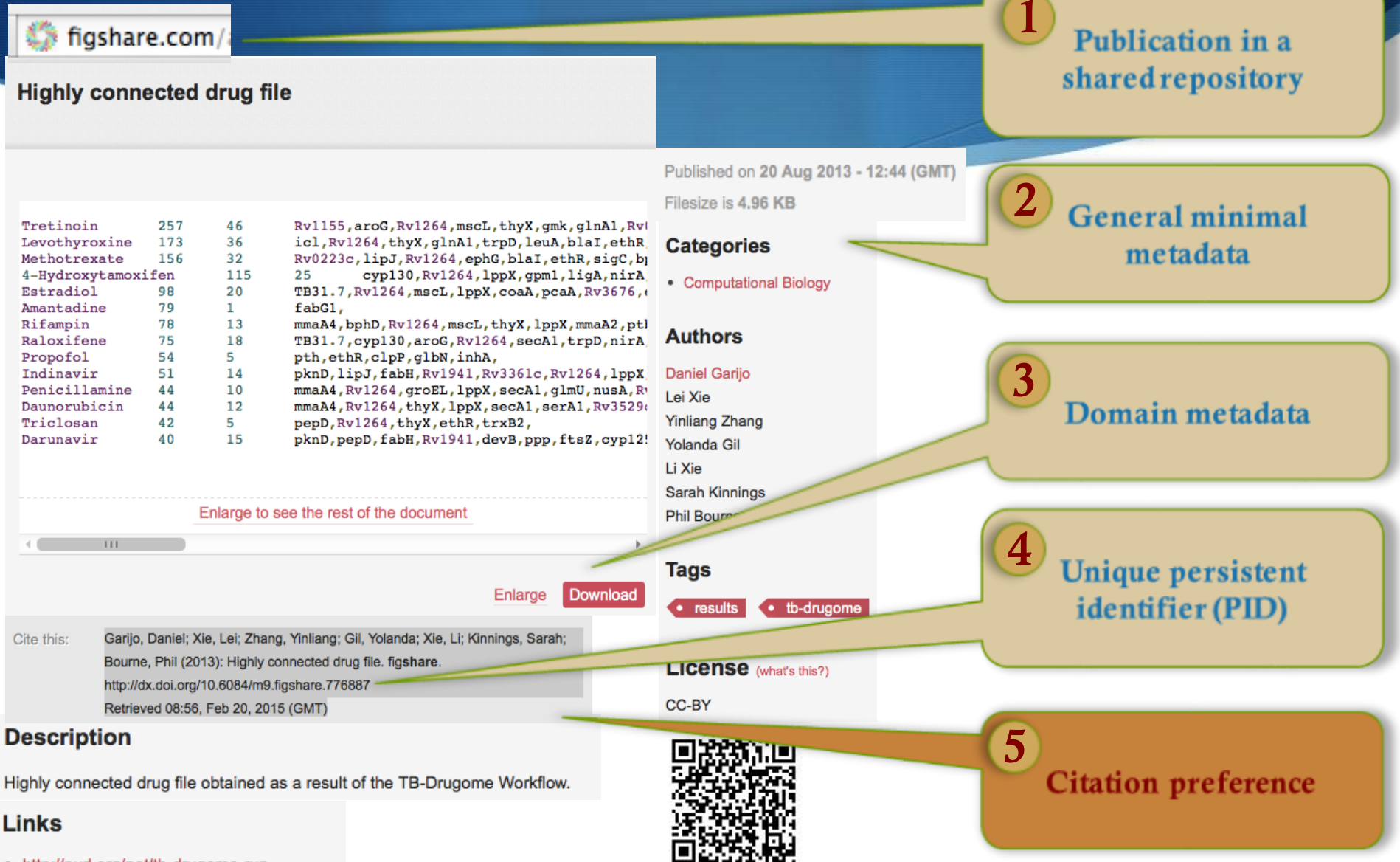
As you may have noticed in the first issue of *PLoS Biology* and again in this issue, there are many places where an alphanumeric string appears after the letters "DOI," such as 10.1371/journal.pbio.0000005 or 10.1371/journal.pbio.0000005.g005. Although some of you may already be acquainted with DOIs, others of you may wonder what they are, how they are used, and why we are using them.

What Are DOIs?

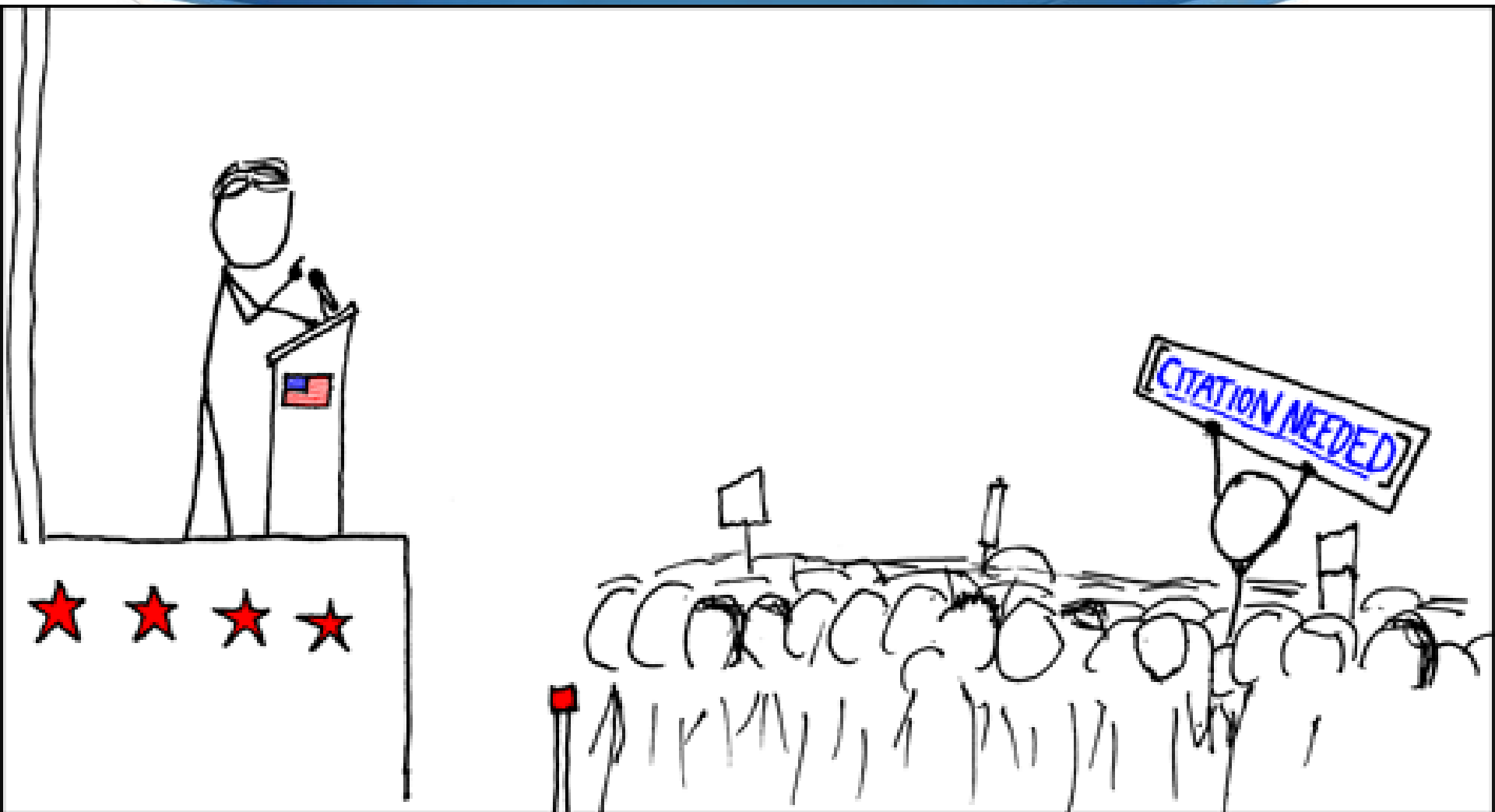
Go to:

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)



Citations: Getting Credit



Citations: Getting Credit

OPEN ACCESS Freely available online

PLoS one

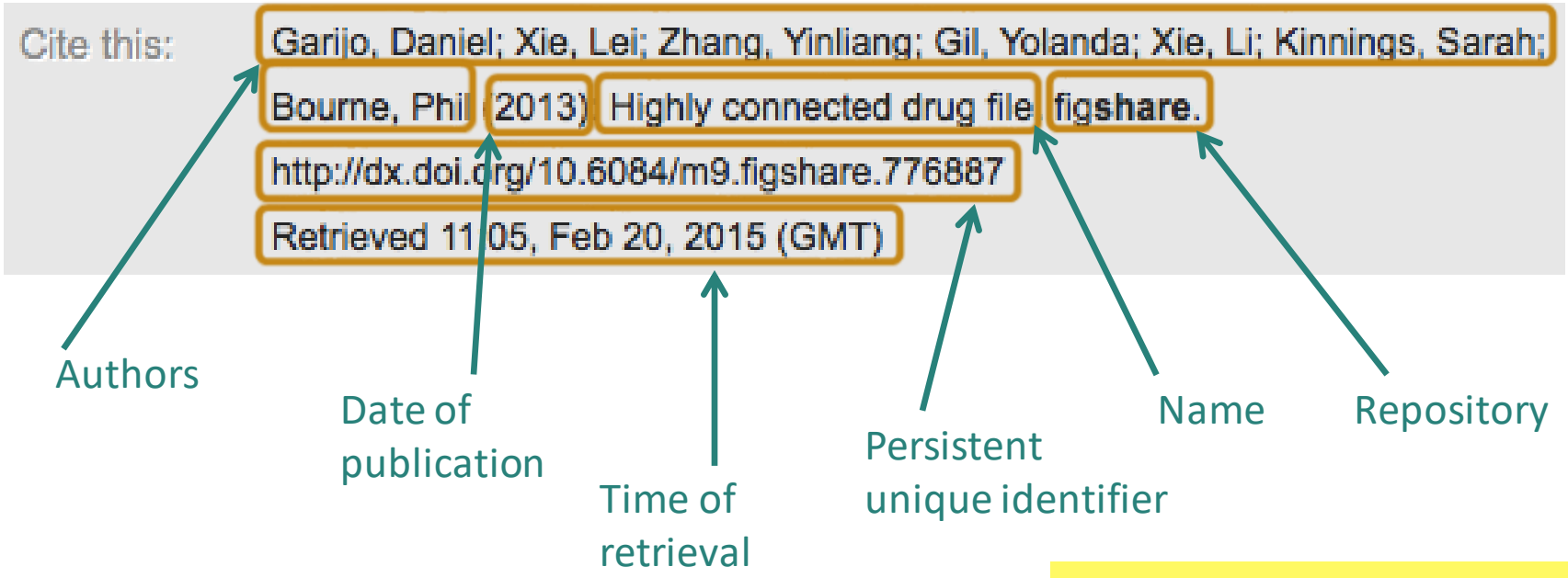
Sharing Detailed Research Data Is Associated with Increased Citation Rate

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

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

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

Data Citation Format



Data repositories and journals often specify how to cite data

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

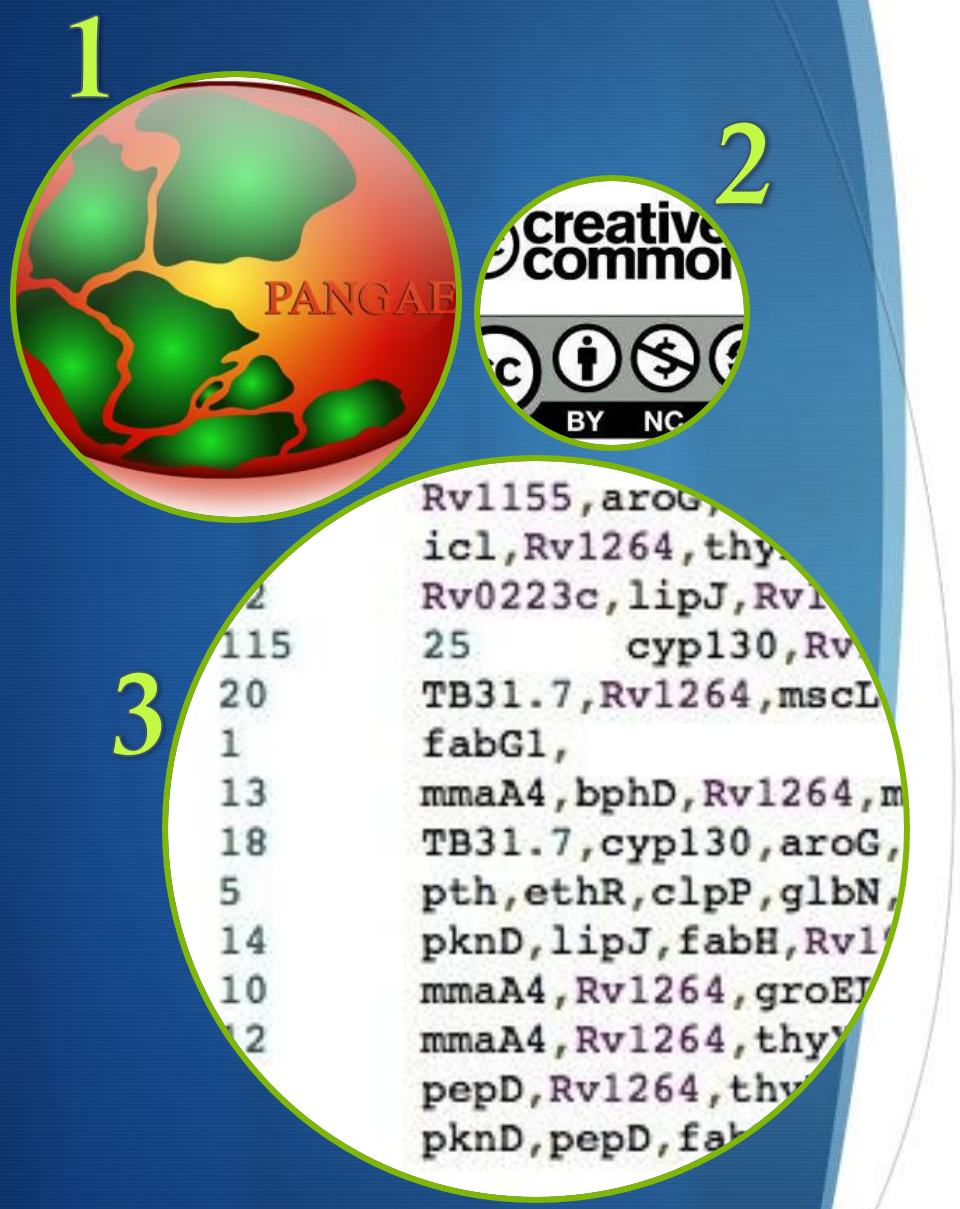
1. Understand what those best practices mean
2. Understand how to implement those best practices

https://www.flickr.com/photos/vizzzual-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

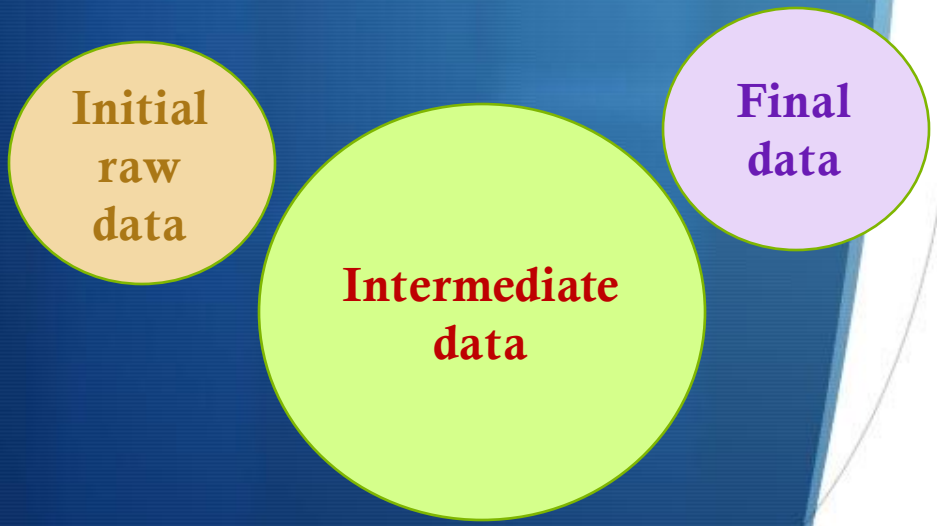


Making Data Accessible: Ideal Approach

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

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. *PRACTICAL EXERCISE*
5. Documenting software with metadata

Part II

1. Documenting software dependencies
2. Documenting methods and workflows
3. Documenting provenance
4. *PRACTICAL EXERCISE*
5. Summary of author checklist

Software in the Scientific Paper of the Future

Part 1.3

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

Onto Soft

NSF ICER-1440323
ICER-1343800

http://www.scientificpaperofthefuture.org

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

The Value of Software

Availability of Software

PLOS | ONE

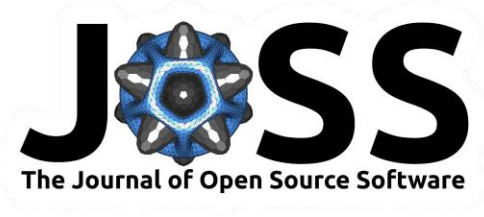
PLOS supports the development of open source software and believes that, for submissions
appropriate open source standards will ensure that the submission conforms to (1) our requi
another researcher can reproduce the experiments described, (2) our aim to promote openn
PLOS journals can be built upon by future researchers. Therefore, if new software or a new a
that the software conforms to the Open Source Definition, have deposited the following three
submission 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.ope
commercial software such as Mathematica and MATLAB does not preclude a paper f
preferred.
- **Documentation for running and installing the software.** For end-user applications
prerequisite; for software libraries, instructions for using the application program inter
- **A test dataset with associated control parameter settings.** Where feasible, result
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 s
for creating user accounts, logging in or otherwise registering personal details. The repositor
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 s

Software Papers and Software Repositories

- Some journal articles describe a piece of software
- Some publications have "software papers" or "software metapapers"

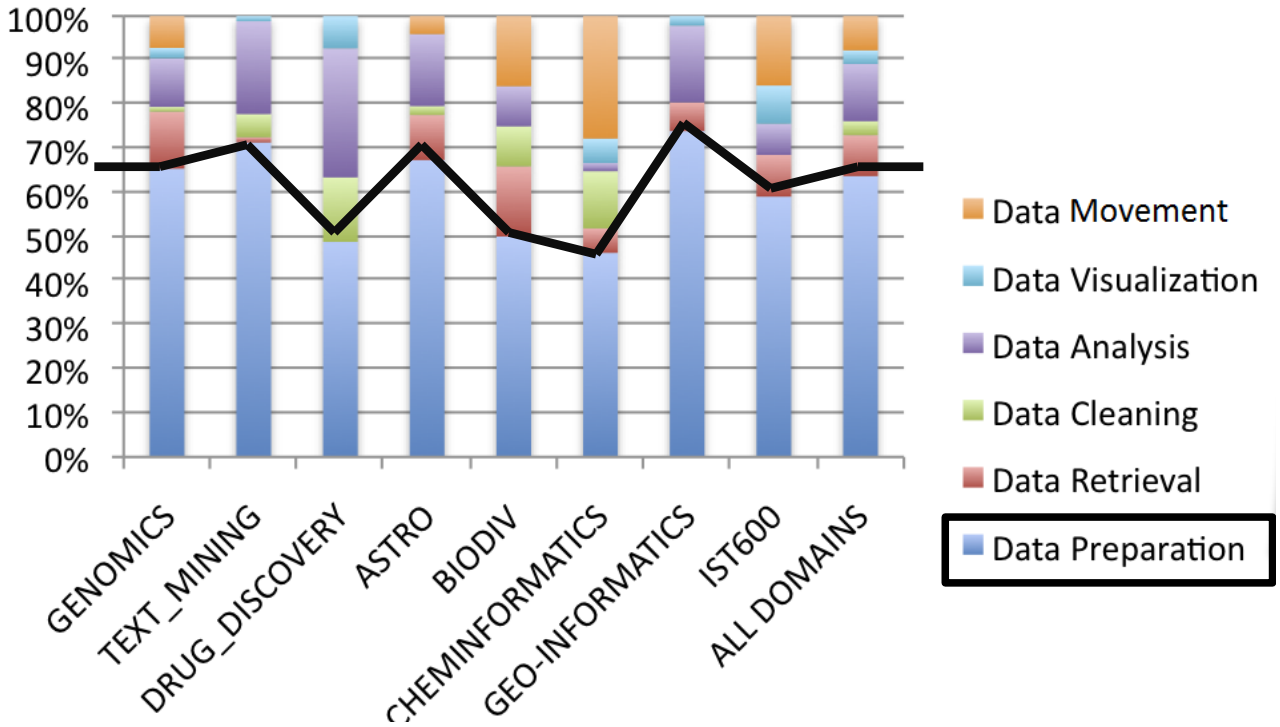


Why Is Scientific Software Not Shared?

- "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!"

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)

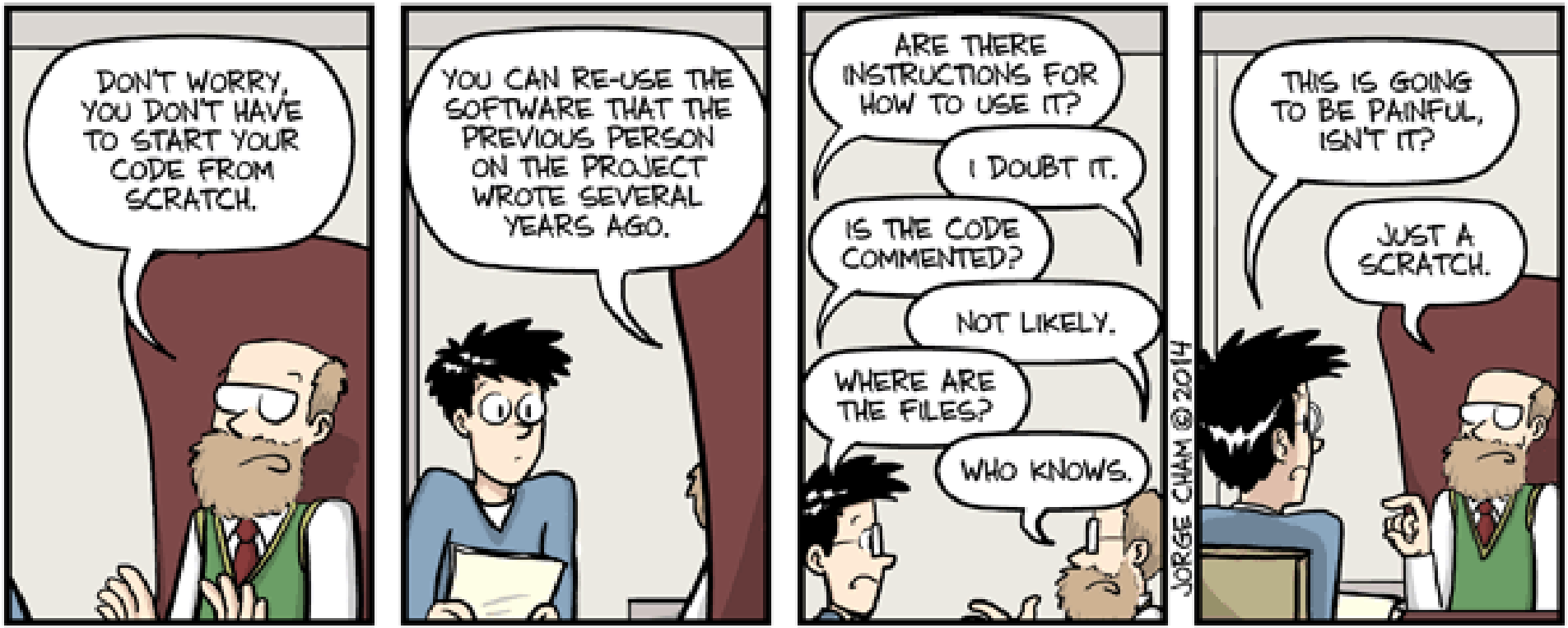


"Common Motifs in Scientific Workflows: An Empirical Analysis." Garijo, D.; Alper, P.; Belhajjame, K.; Corcho, O.; Gil, Y.; and Goble, C. Future Generation Computer Systems, 2013.

“Dark Software”

- Models that are not published
  - E.g. from a PhD thesis
- Data preparation software
- Visualization software

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


DON'T WORRY, YOU DON'T HAVE TO START YOUR CODE FROM SCRATCH.
YOU CAN RE-USE THE SOFTWARE THAT THE PREVIOUS PERSON ON THE PROJECT WROTE SEVERAL YEARS AGO.
ARE THERE INSTRUCTIONS FOR HOW TO USE IT?
I DOUBT IT.
IS THE CODE COMMENTED?
NOT LIKELY.
WHERE ARE THE FILES?
WHO KNOWS.
JORGE CHAM © 2014
THIS IS GOING TO BE PAINFUL, ISN'T IT?
JUST A SCRATCH.
WWW.PHDCOMICS.COM

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/vizzzual-dot-com/2655969483/

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

Goals of this Section

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

Best Practices

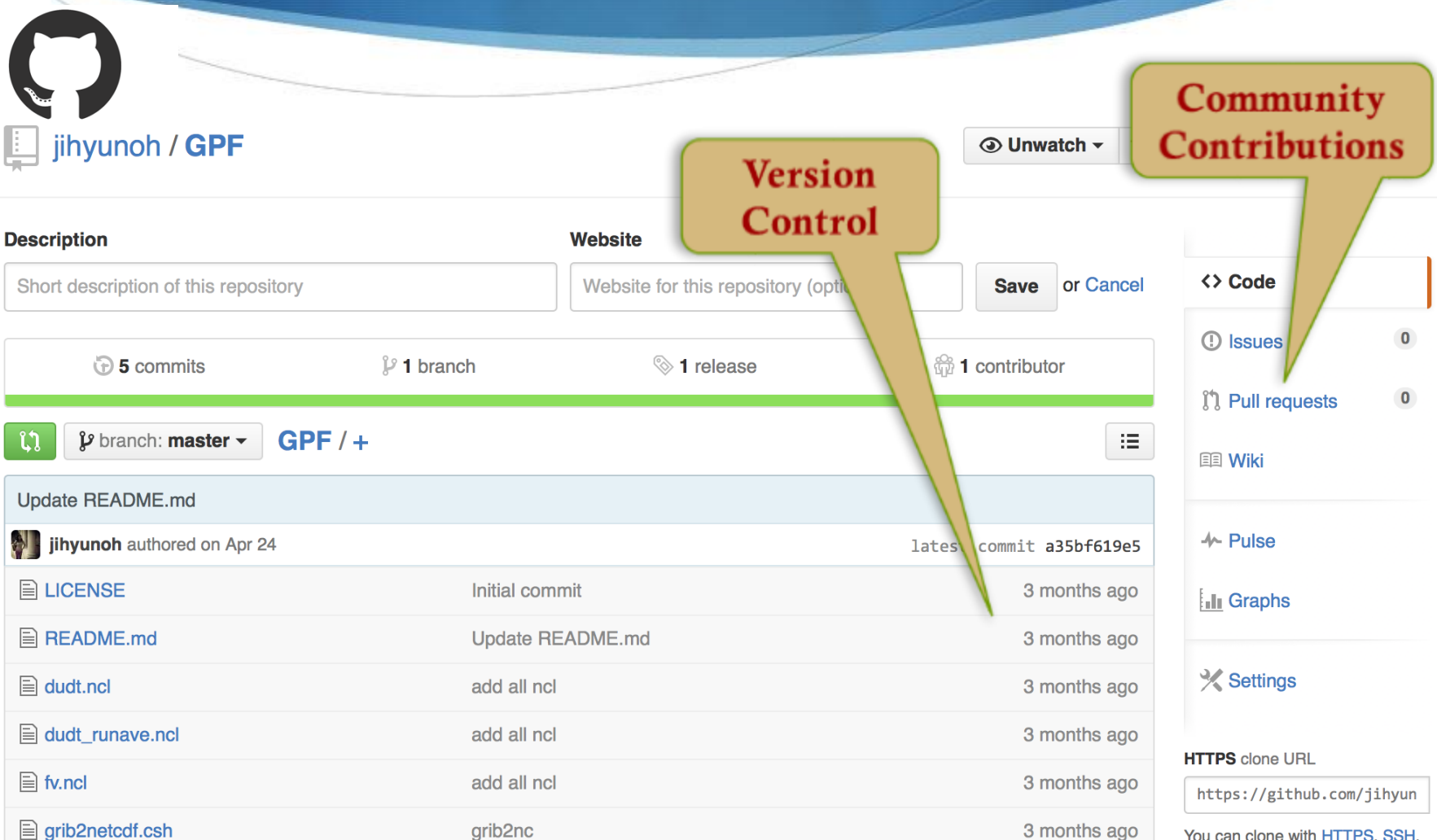
1. Accessible from a public location
2. License
3. Citation

Making Software Accessible from a Public Location

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** (e.g., Zenodo), treating code like data
  - Very easy and simple
  - It allows you to get a DOI
- **Use a code repository** (e.g., GitHub, BitBucket)
  - Beneficial if you have other users or want to track new versions
  - Some will give you a DOI (e.g., GitHub)
- **Create a formal community project** (e.g., in Apache)
  - Very involved, but very beneficial if you have many users

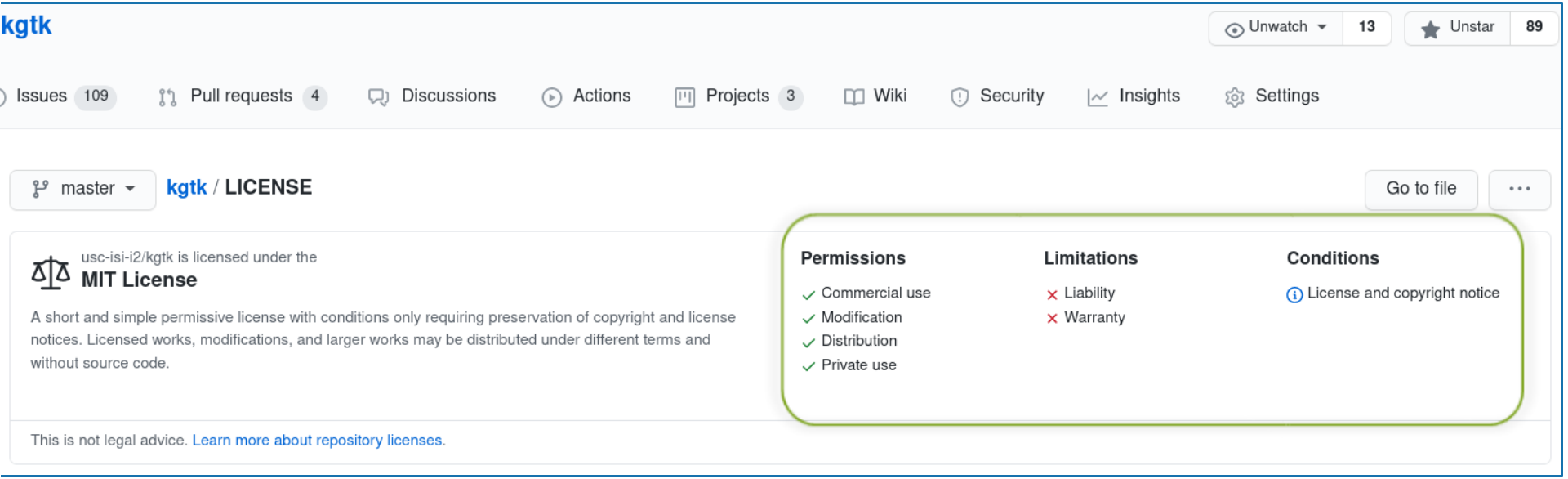
Publishing Software in a Code Repository



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
  - "Copyleft" license (ex: GNU General Public License (GPL))
  - "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

Some repositories can help you choose a license



Software Citation

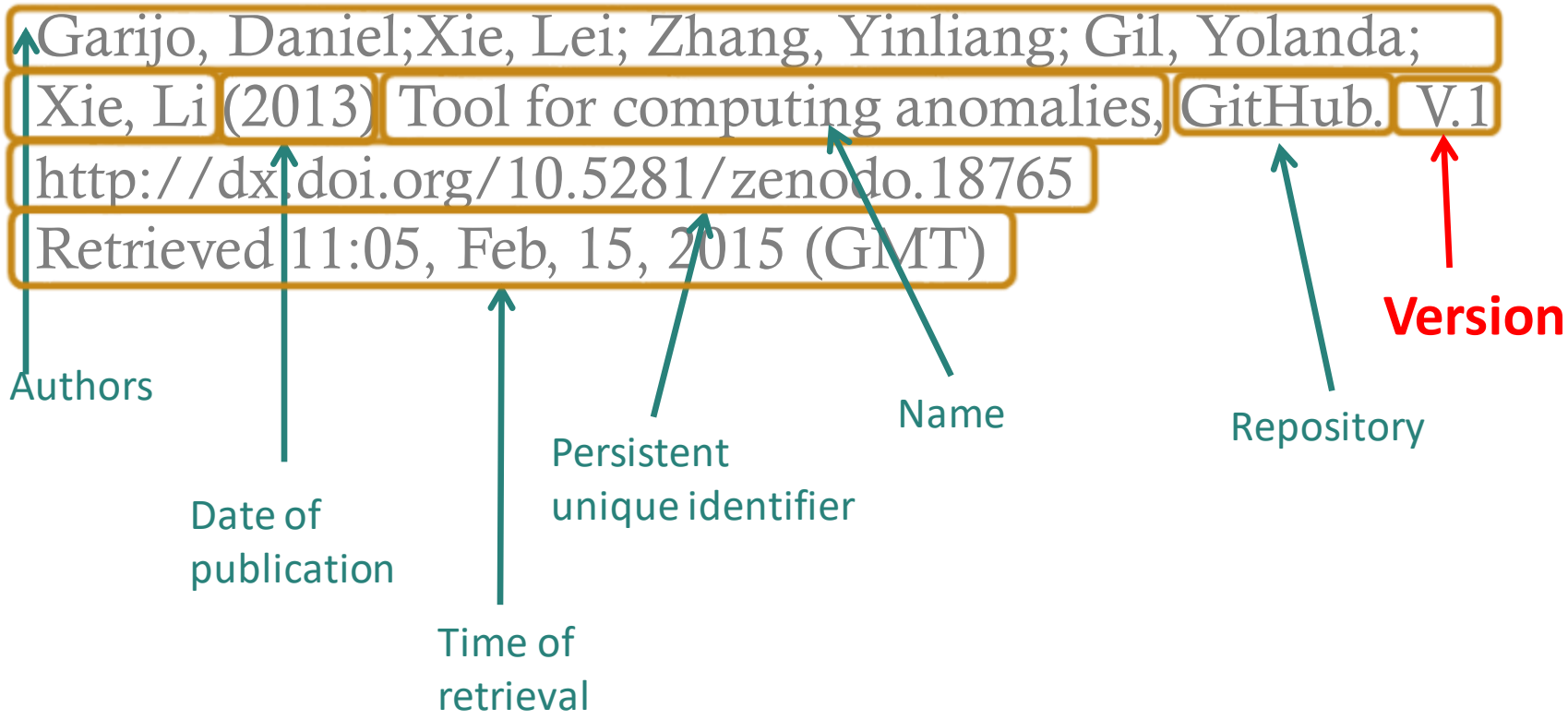
- What do you want to cite?
  - Code? Project Website? Commit? Release?
- Use a persistent unique identifier (PURL or DOI)
  - Analogous to identifiers for data
- Software sharing repositories are beginning to offer the ability to assign DOIs

More information: Smith AM, Katz DS, Niemeyer KE, FORCE11 Software Citation Working Group. (2016) Software Citation Principles. *PeerJ Computer Science* 2:e86.
DOI: 10.7717/peerj-cs.86

https://citation-file-format.github.io/ (recently adopted by GitHub)

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

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

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

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

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. Ma[illegible]g data accessible
3. [illegible]king software [illegible]cessible
4. *PRACTICAL EXERCISE*
5. Documenting software with metadata

Part II

1. Documenting software dependencies
2. Documenting methods and workflows
3. Documenting provenance
4. *PRACTICAL EXERCISE*
5. Summary of author checklist

Practical Exercise: Obtain a DOI for your software (1)

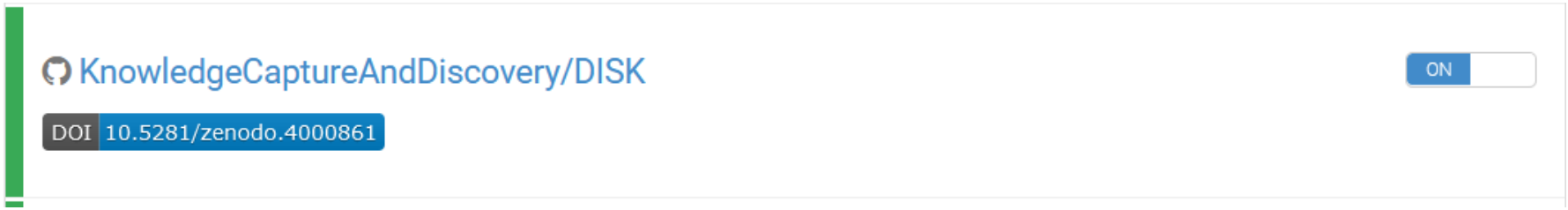
- Use your GitHub credentials to log into **Zenodo** (https://zenodo.org)
- **Authorize** Zenodo to access your GitHub account
- In **settings** -> GitHub, your repository should appear accessible
- Flip the switch to "**ON**"
- More details at https://guides.github.com/activities/citable-code/

Practical Exercise: Obtain a DOI for your software (2)

- Add code to your GitHub repository. When you are ready, click on "releases" and select "Create new release".
- Describe your release
- Give it a proper **version number**!

Semantic versioning: https://semver.org/

- Now go to your Zenodo page. If everything went correctly, you should see a DOI for your GitHub repository
- Now you can copy the blue Zenodo badge with the DOI back in your GitHub readme file



Documenting Software through Metadata

Part 1.5

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

SCIENTIFIC PAPER OF THE FUTURE — Onto Soft

Onto Soft

NSF ICER-1440323 ICER-1343800

http://www.scientificpaperofthefuture.org

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

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. *PRACTICAL EXERCISE*
5. Documenting software with metadata

Part II

1. Documenting software dependencies
2. Documenting methods and workflows
3. Documenting provenance
4. *PRACTICAL EXERCISE*
5. Summary of author checklist

Software Repositories

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

CRAN

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

A vocabulary for describing software: Codemeta

Property	Type	Description
softwareSuggestions	SoftwareSourceCode	Optional dependencies , e.g. for optional features, code development, etc.
maintainer	Person	Individual responsible for maintaining the software (usually includes an email contact address)
contIntegration	URL	link to continuous integration service
buildInstructions	URL	link to installation instructions/documentation
developmentStatus	Text	Description of development status, e.g. Active, inactive, suspended. See repostatus.org
embargoDate	Date	Software may be embargoed from public access until a specified date (e.g. pending publication, 1 year from publication)
funding	Text	Funding source (e.g. specific grant)
issueTracker	URL	link to software bug reporting or issue tracking system
referencePublication	ScholarlyArticle	An academic publication related to the software.
readme	URL	link to software Readme file

- Schema.org extension (findable by search engines)

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…"

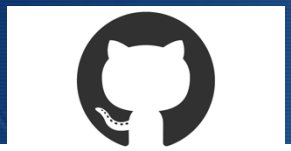
What if…

- **… 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



jihyunoh / GPF

Unwatch 1 | Star 0 | Fork 0

Description

Short description of this repository

Website

Website for this repository (optional)

Save or Cancel

5 commits | 1 branch | 1 release | 1 contributor

branch: master | GPF / +

Update README.md

jihyunoh authored on Apr 24 — latest commit a35bf619e5

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
pgf.ncl	add all ncl	3 months ago
plot_pgf_x.ncl	add all ncl	3 months ago
plot_ududx_runave.ncl	add all ncl	3 months ago

Code

Issues 0

Pull requests 0

Wiki

Pulse

Graphs

Settings

HTTPS clone URL

https://github.com/jihyun

You can clone with HTTPS, SSH, or Subversion.

Clone in Desktop

Download ZIP

Describing Software with Codemeta

CodeMeta generator

Most fields are optional. Mandatory fields will be highlighted when generating Codemeta.

The software itself

Name
My Software
the software title

Description
My Software computes ephemerides and orbit propagation. It has been developed from early '80.

Creation date
YYYY-MM-DD

First release date
YYYY-MM-DD

License
from SPDX licence list

Discoverability and citation

Unique identifier
10.151.xxxxx
such as ISBNs, GTIN codes, UUIDs etc.. http://schema.org/identifier

Application category
Astronomy

Keywords
ephemerides, orbit, astronomy

Funding
PRA_2018_73
grant funding software development

Funder
Università di Pisa
organization funding software development

Authors and contributors can be added below

Development community / tools

Code repository
git+https://github.com/You/RepoName.git

Continuous integration
https://travis-ci.org/You/RepoName

Issue tracker
https://github.com/You/RepoName/issues

Related links

Run-time environment

Programming Language
C#, Java, Python 3

Runtime Platform
.NET, JVM

Operating System

Current version of the software

Version number
1.0.0

Release date
YYYY-MM-DD

Download URL

Additional Info

Reference Publication
https://doi.org/10.1000/xyz123

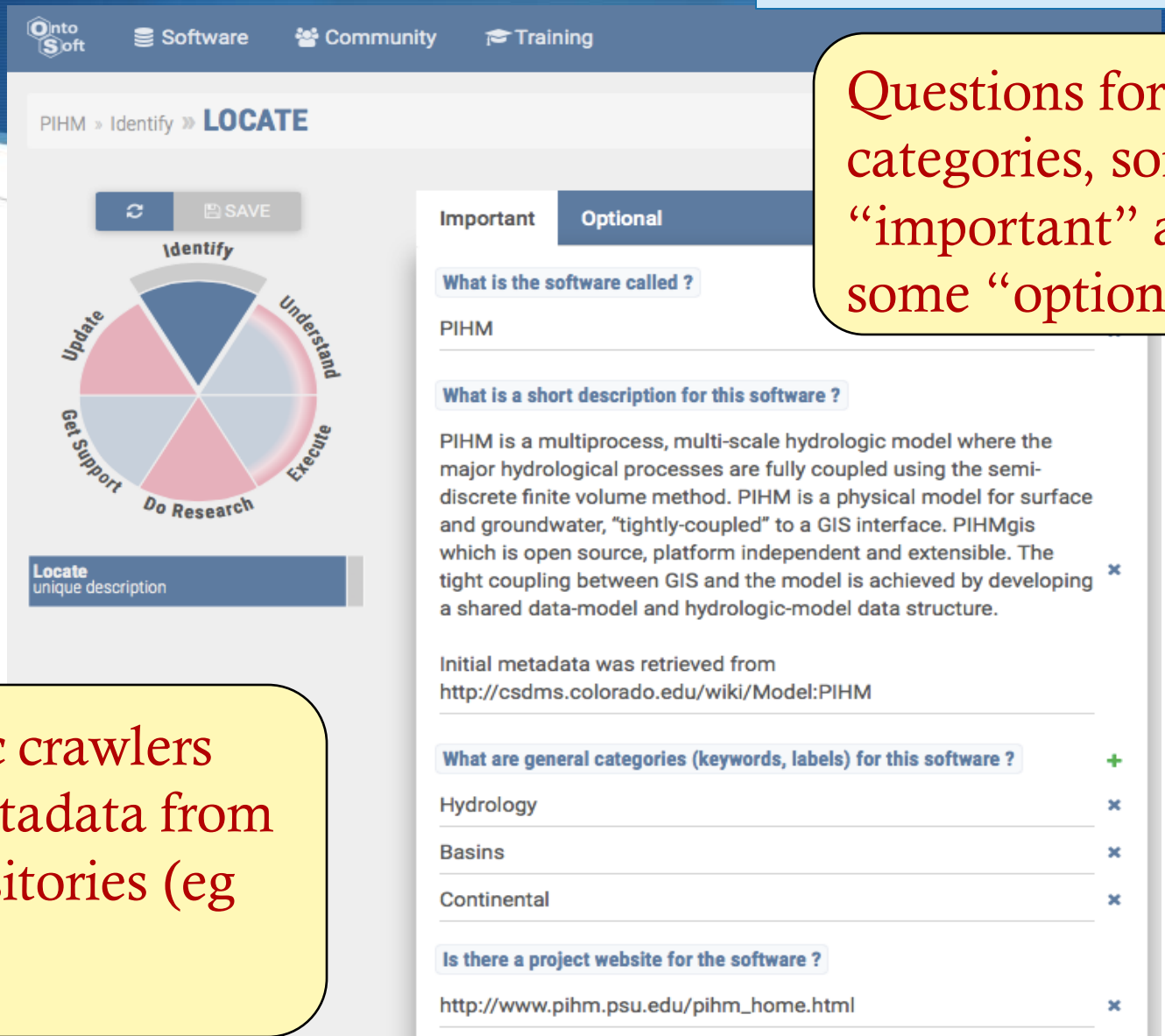
Development Status

see www.repostatus.org for details

https://codemeta.github.io/codemeta-generator/ (manually) or https://somef.readthedocs.io/en/latest/ (automatically)

Describing Software with OntoSoft

http://www.ontosoft.org/portal

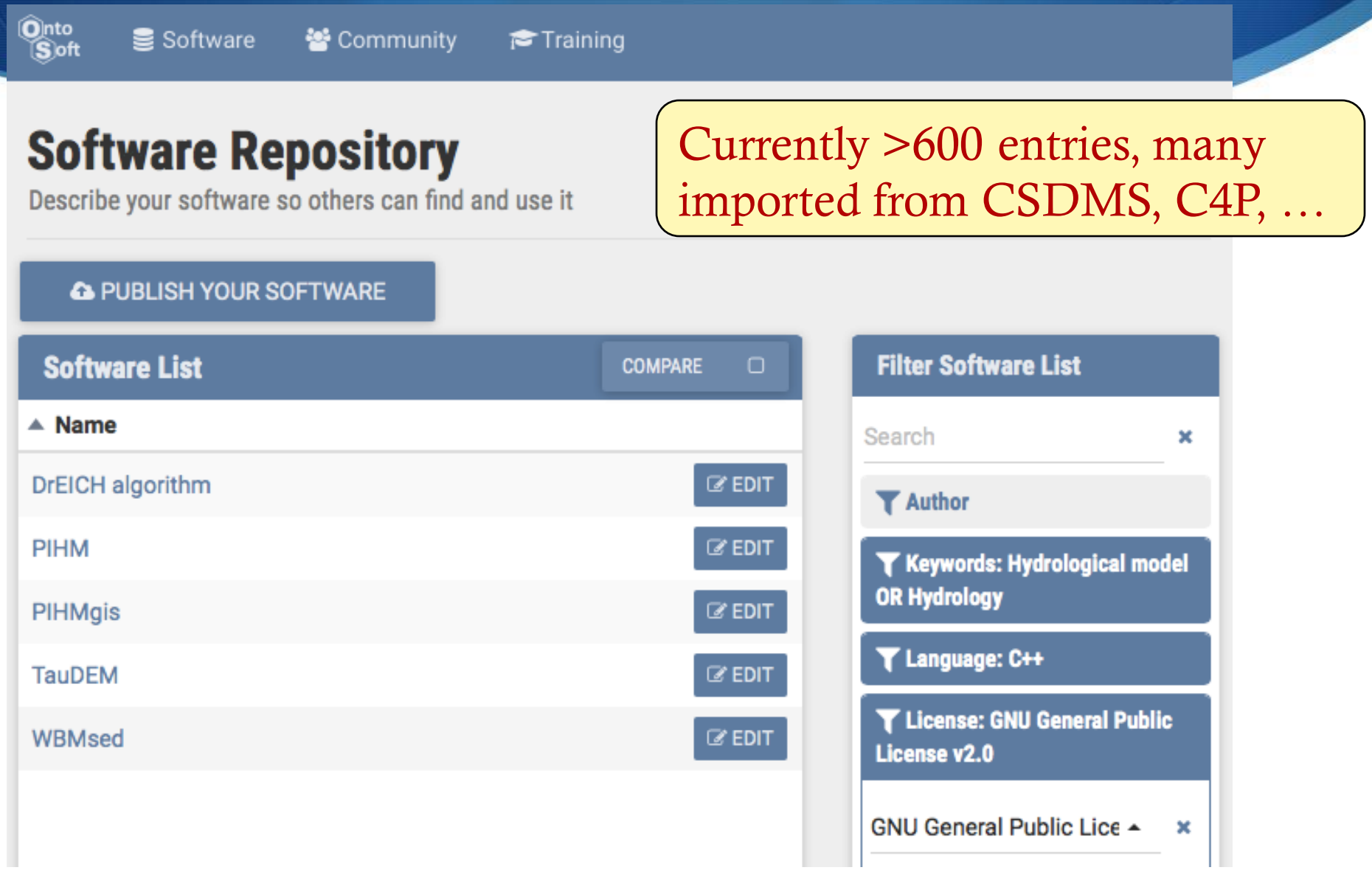


Questions for 6 top categories, some "important" and some "optional"

Automatic crawlers import metadata from code repositories (eg GitHub)

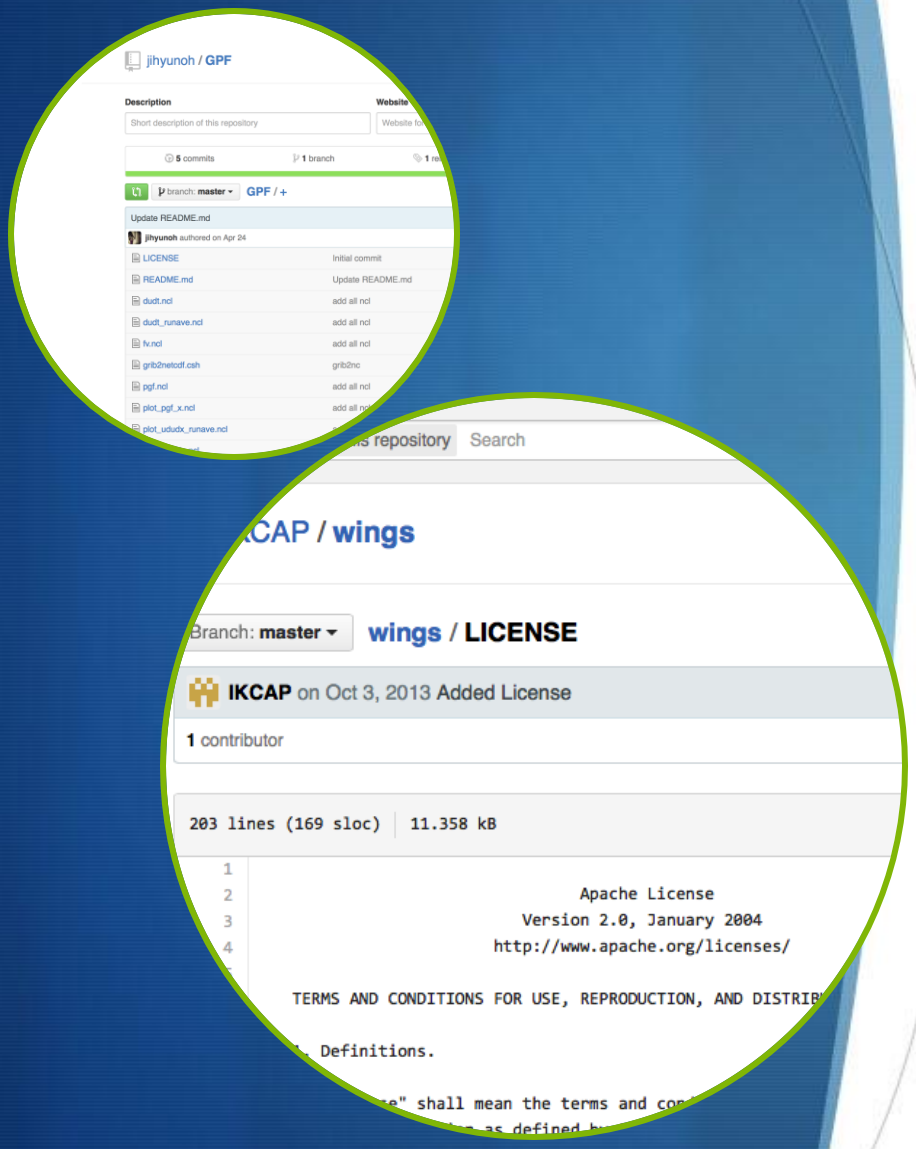
Finding Software with OntoSoft

http://www.ontosoft.org/portal

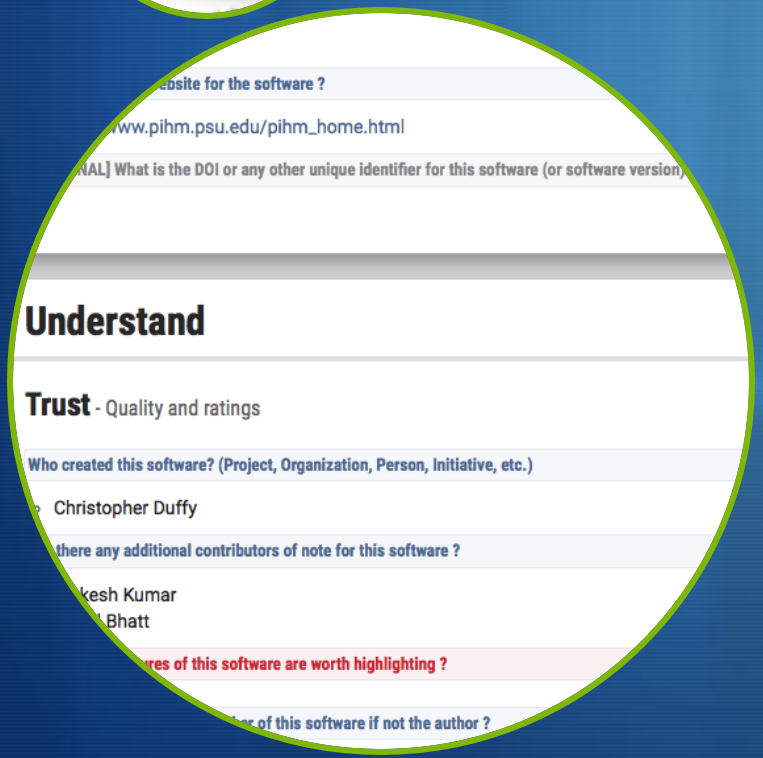


Currently >600 entries, many imported from CSDMS, C4P, …

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.
2. Save the metadata as HTML, XML,…
   - Use codemeta generator to create a Codemeta file
3. Post the metadata on your code site

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. *PRACTICAL EXERCISE*
5. Documenting software with metadata

Part II

1. Describing software dependencies
2. Documenting methods and workflows
3. Documenting provenance
4. *PRACTICAL EXERCISE*
5. Summary of author checklist

Describing software dependencies

Part 2.1

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

NSF ICER-1440323
ICER-1343800

http://www.scientificpaperofthefuture.org

Software dependency hell

"Oh, you can't run my code? But it works in my machine…"

- Package dependencies may have incompatibilities
  - E.g., some dependencies may work in Python 3.6 but do not in Python 3.7…
- Some libraries may require different versions installed
  - In one project, library A requires numpy=1.0, but in my laptop I installed numpy 2.0 for project B
- Different operative systems may support different libraries

How to keep track of your software dependencies?

Virtual environments

```
PS C:\Users\dgarijo\Documents\GitHub\SM2KG> .\env\Scripts\activate
(env) PS C:\Users\dgarijo\Documents\GitHub\SM2KG> python --version
Python 3.7.7
(env) PS C:\Users\dgarijo\Documents\GitHub\SM2KG> deactivate
PS C:\Users\dgarijo\Documents\GitHub\SM2KG> python --version
Program 'python' failed to run: No application is associated with the
+ python --version
```

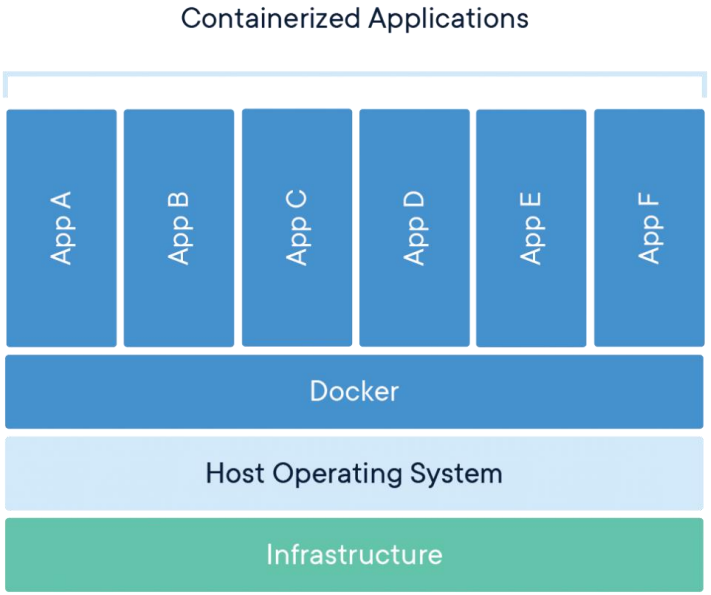
Containers

Package managers

Virtual machines

Software Containers

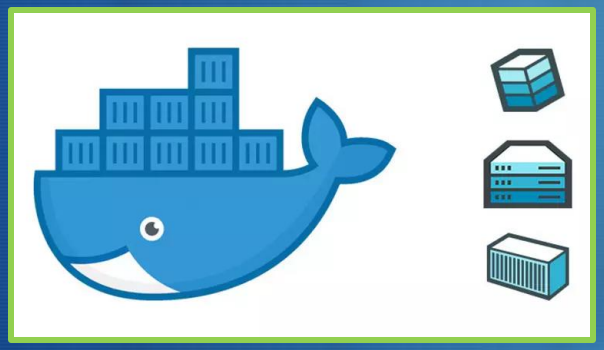
- Track the software dependencies and OS
- Software image: executable which specifies a full
  - computational environment
  - code, runtime, system tools, system libraries and settings
- Container: virtualized computational environment
- Used to run one or multiple software images



https://www.docker.com/resources/what-container

Documenting Software dependencies: Simplest Approach

1. Keep track of your dependencies
    1. Describe precisely the requirements in your readme
    2. Preserve your environment (requirements.txt., pom.xml, etc.)

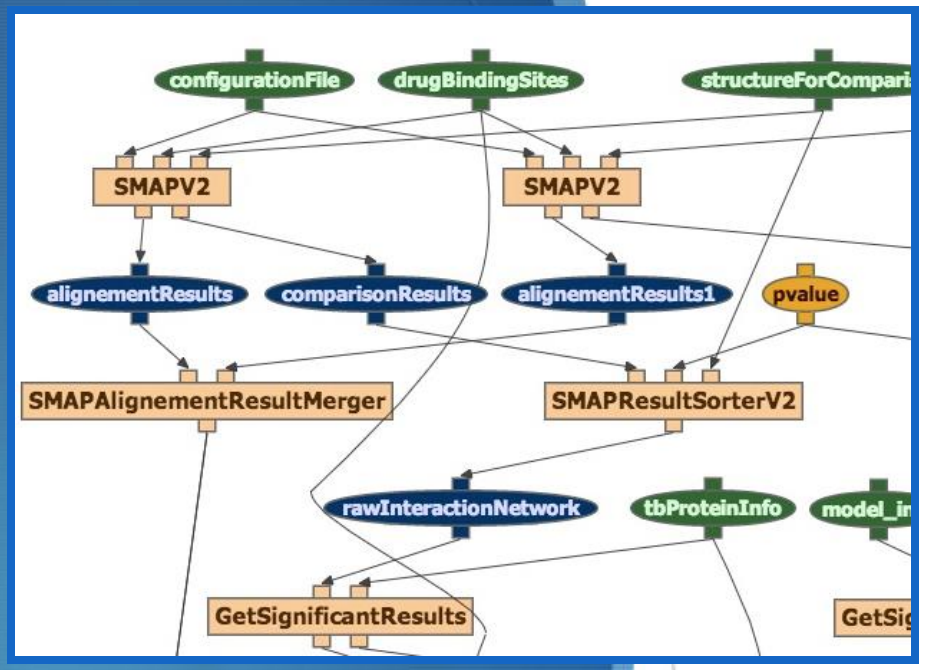


Ideal Approach

1. Generate one or multiple DockerFiles
   - E.g., develop version, main version, etc.
2. Make image available in an image repository
   - E.g., DockerHub (there are others)
3. Describe your image with basic metadata

https://www.icm.es/2019/12/07/docker-containers/

Methods and Workflows in the Scientific Paper of the Future



Part 2.2

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

http://www.scientificpaperofthefuture.org

ICER-1440323
ICER-1343800

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. *PRACTICAL EXERCISE*
5. Documenting software with metadata

Part II

1. Documenting software dependencies
2. Documenting methods and workflows
3. Documenting provenance
4. *PRACTICAL EXERCISE*
5. Summary of author checklist

Methods Described in Text Are Ambigous and Incomplete

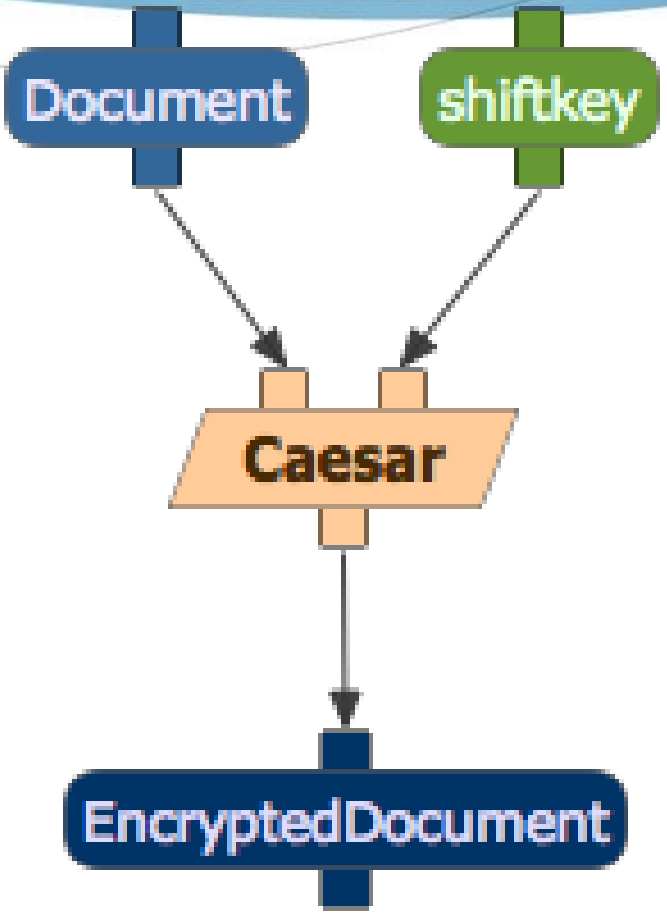
- "**Ambiguity** in program descriptions leads to the possibility, if not the certainty, that a given natural language description can be converted into computer code in various ways, each of which may lead to different numerical outcomes." [Ince et al 2012]
- Analysis of 18 quantitative papers published 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]

Goals of this Section

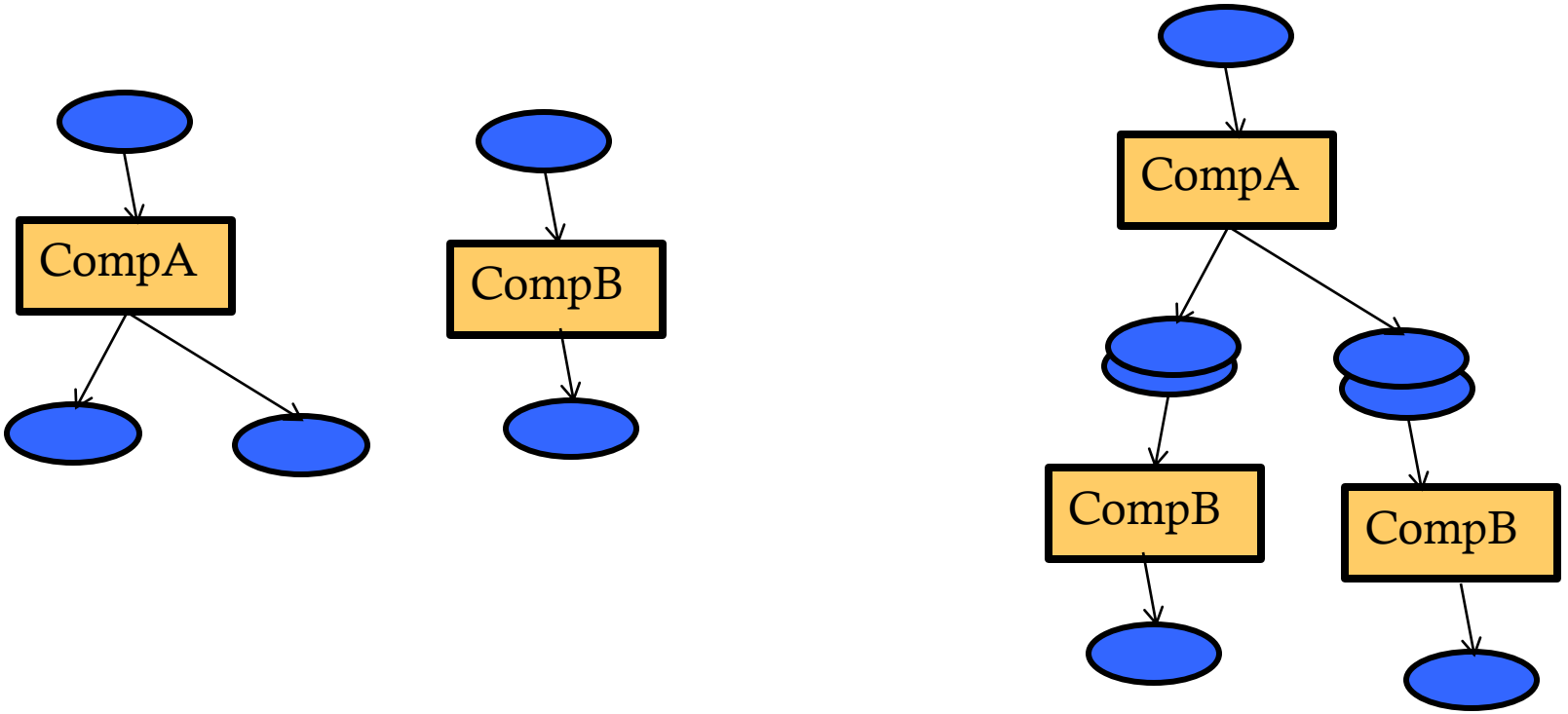
1. Understand what are methods and provenance is in a scientific article
2. Understand how to document methods and provenance properly in an article

https://www.flickr.com/photos/vizzzual-dot-com/2655969483/

Programs as Black Boxes: Functions with Inputs, Outputs, and Parameters

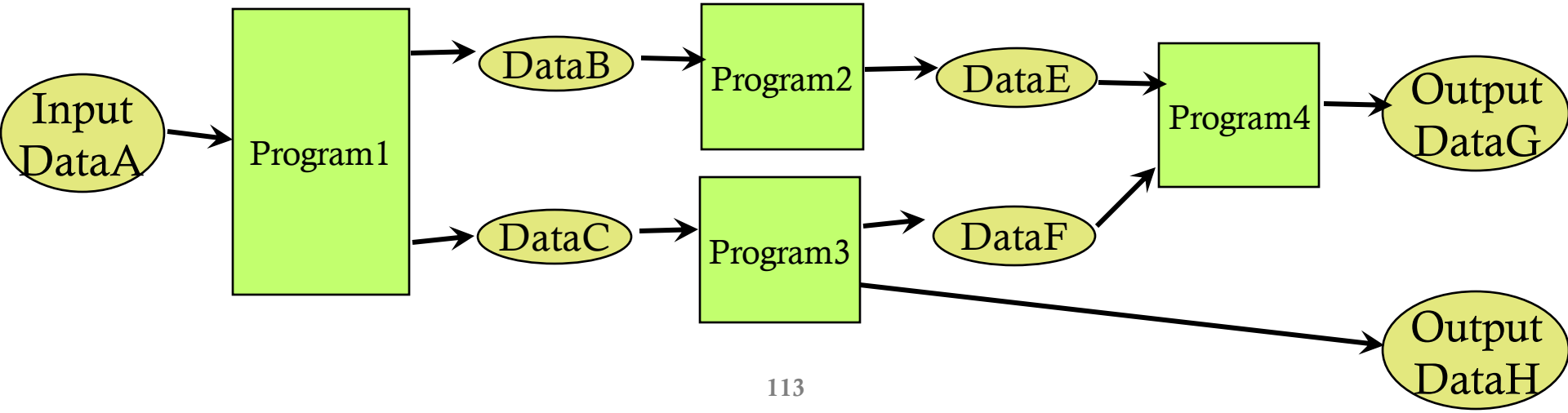


Composing Functions



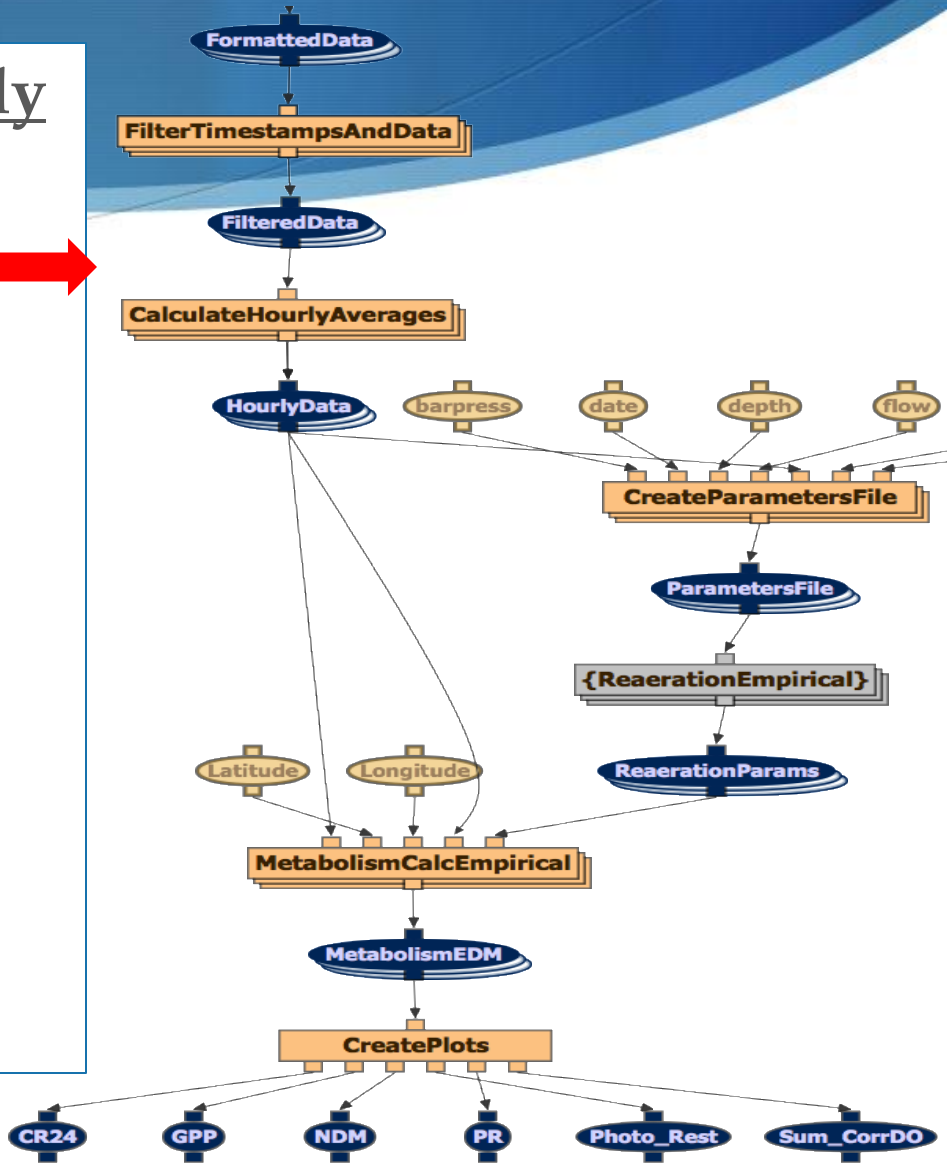
Computational Workflows

- Workflow is represented as a graph of connected nodes
  - Nodes represent programs and data (alternatively)
  - Links represent how data flows from program to program (output to input)
- Computational workflows are compositions of programs
  - No user interaction during execution
  - No cycles allowed!

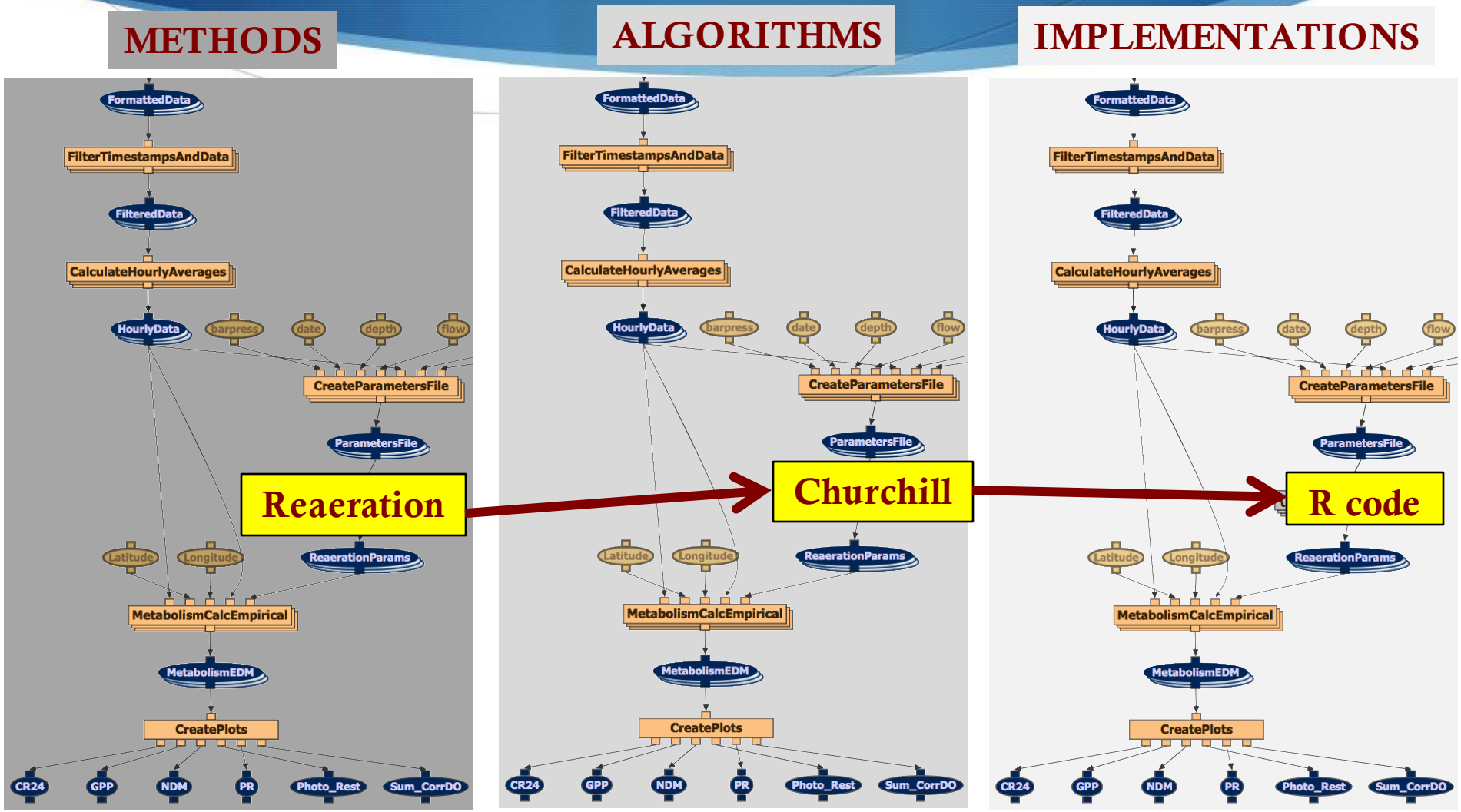


Workflows as Representations of Computational Methods

- Computational workflow **only** contains computational steps
  - E.g., water metabolism
- Workflows can include manual steps
  - E.g., creating a figure, cleaning data
- Workflows may access web services
  - E.g., access databases in biology

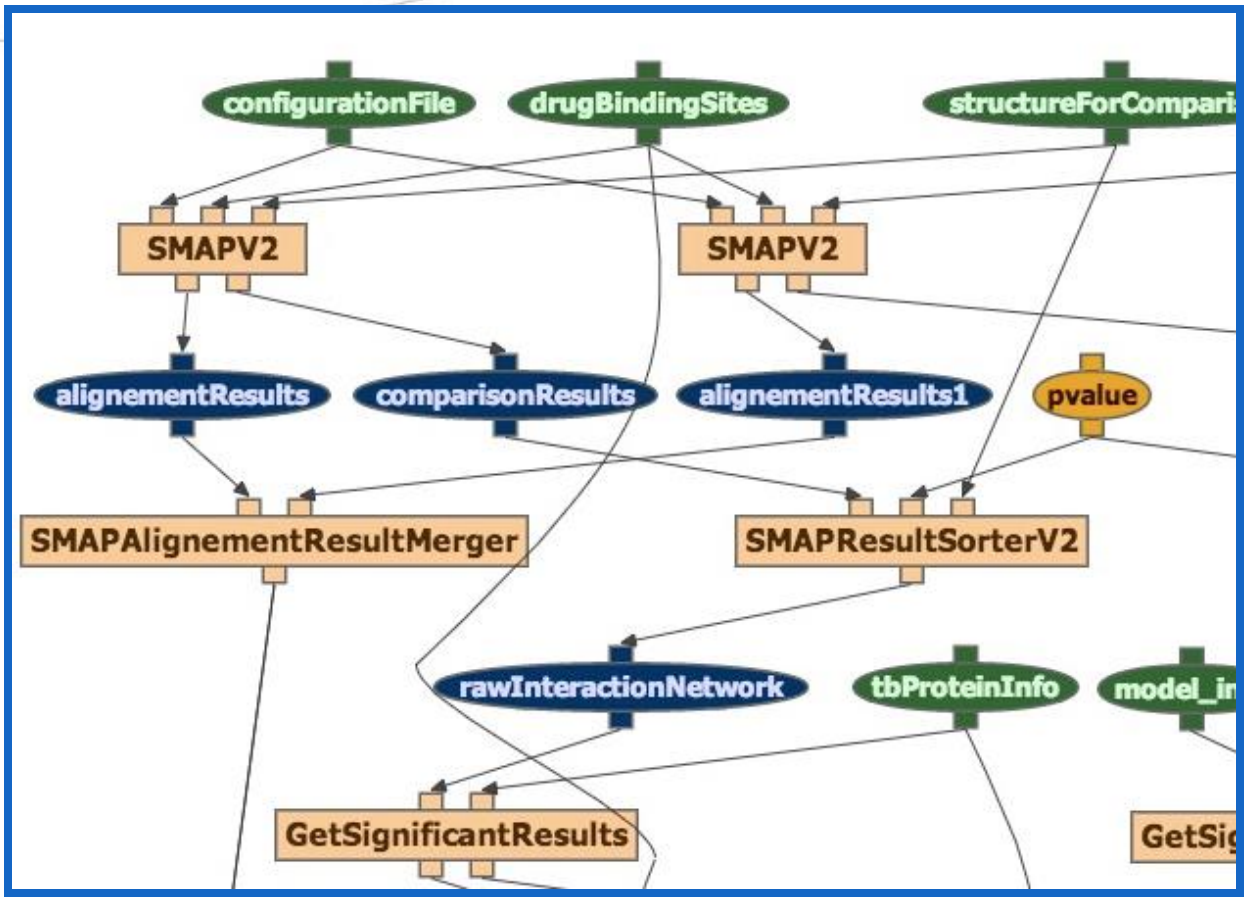


Describing a Method at Different Levels of Abstraction



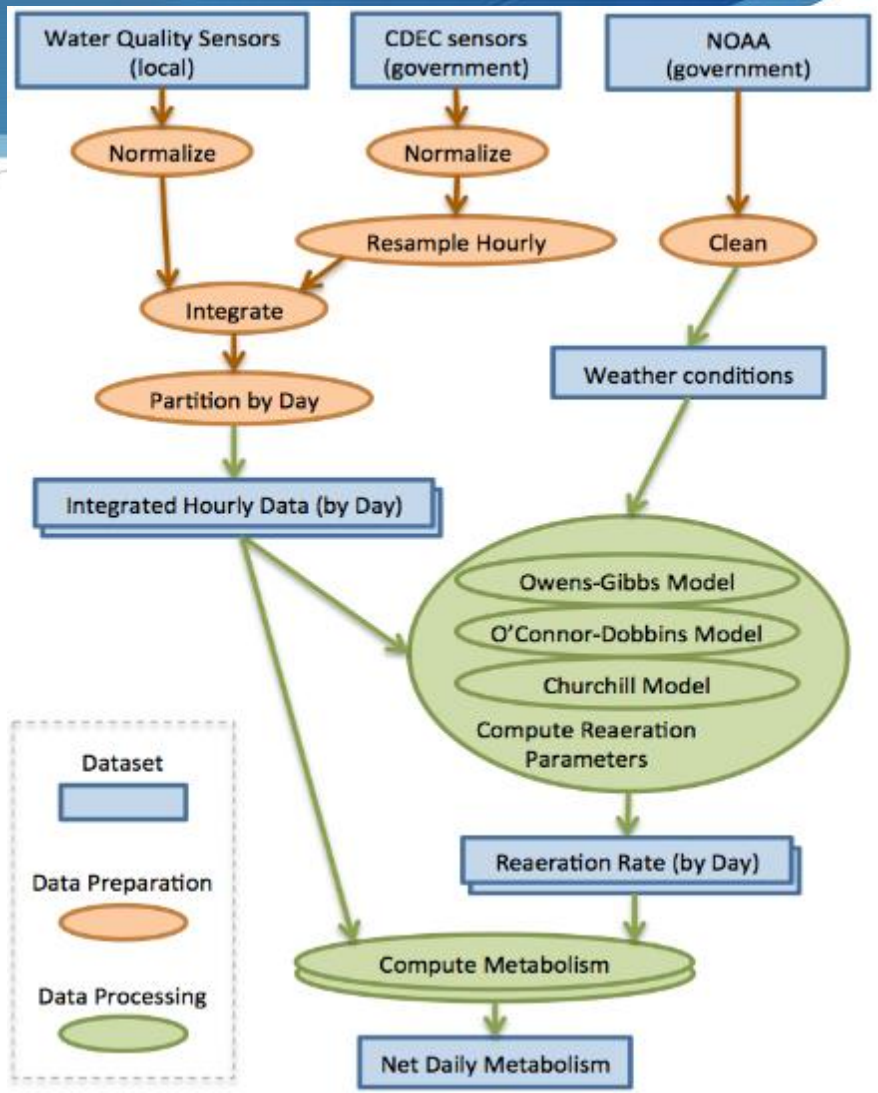
What the Paper Says Versus What the Actual Software Does

(from [Garijo et al 2013])

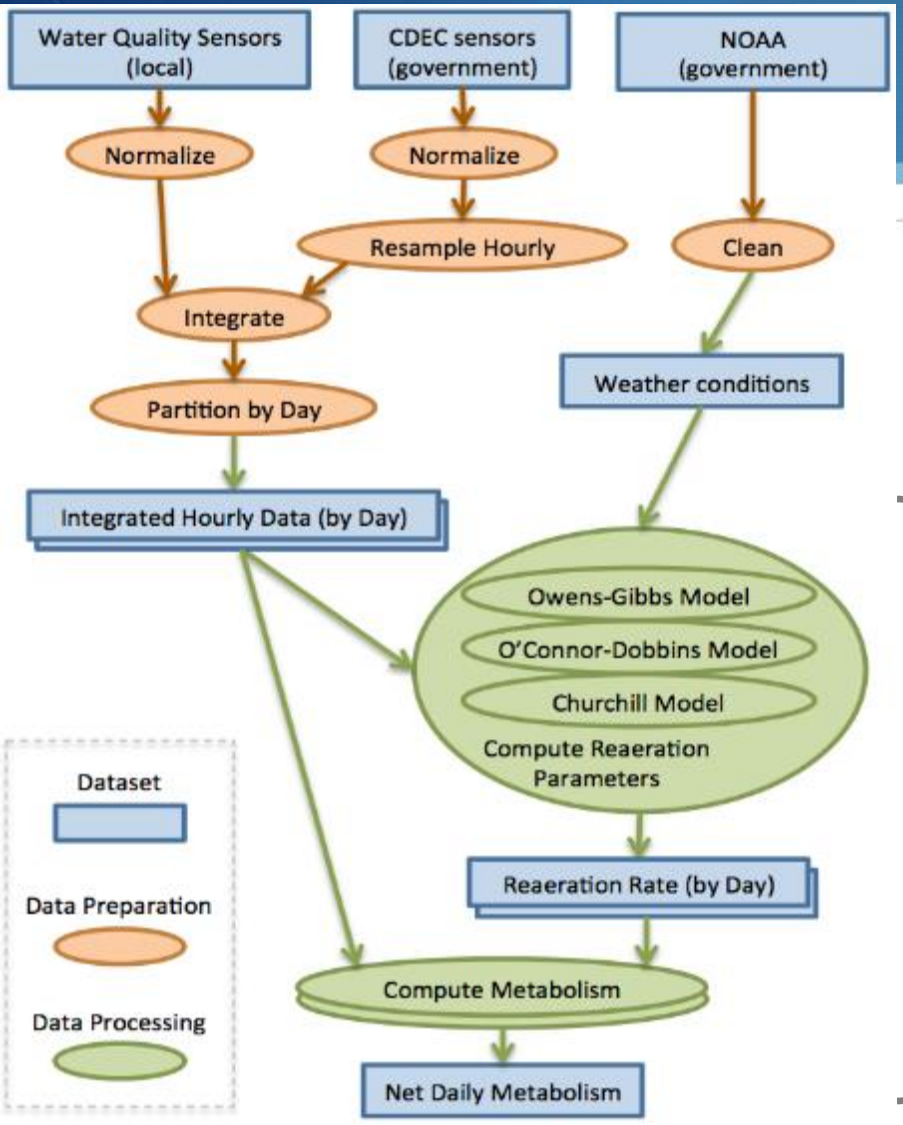


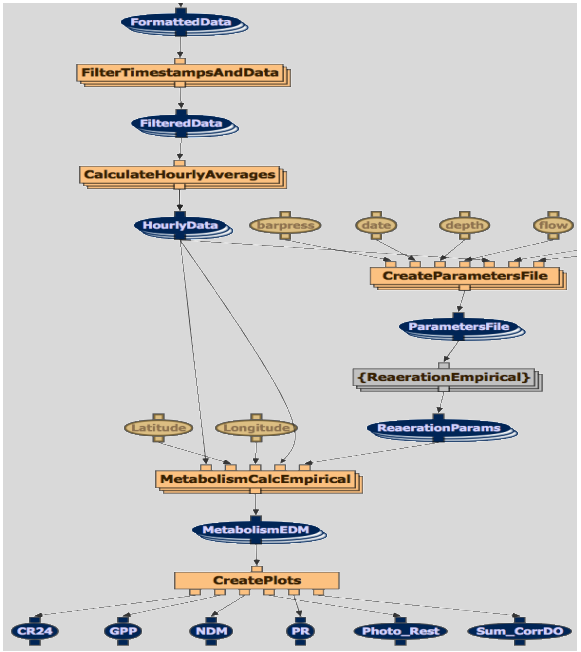
Developing Workflows: How to Sketch a Workflow

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

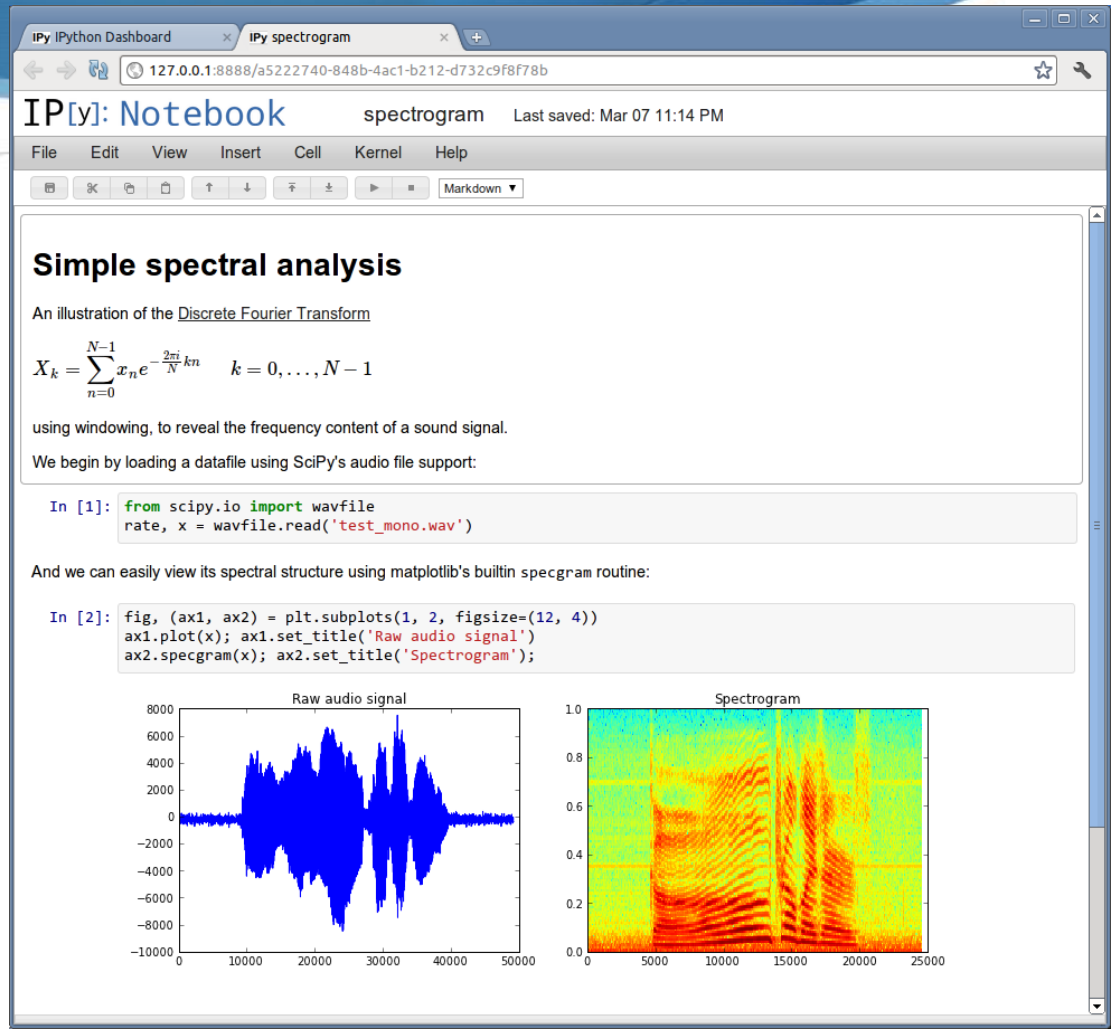
jupyter

Sweave = R · LaTeX

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

MADAGASCAR

http://ipython.org/notebook.html



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. *PRACTICAL EXERCISE*
5. Documenting software with metadata

Part II

1. Documenting software dependencies
2. Documenting methods and workflows
3. Documenting provenance
4. *PRACTICAL EXERCISE*
5. Summary of author checklist

Provenance in the Scientific Paper of the Future

Part 2.3

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

http://www.scientificpaperofthefuture.org

ICER-1440323
ICER-1343800

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

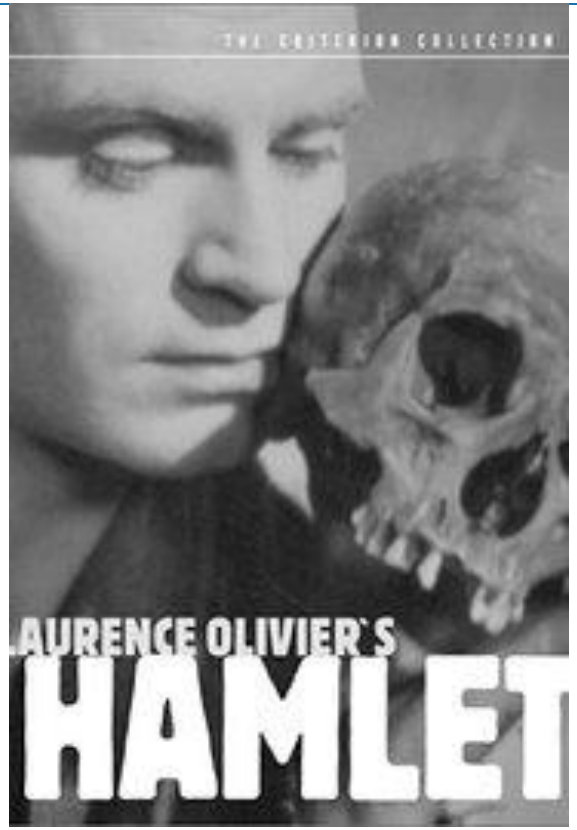
https://www.flickr.com/photos/alterschwede08/3203630740/ (CC BY-ND 2.0)

The Many Meanings of Provenance

- A signature

- A method

- A document



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)
http://www.imdb.com/title/tt0040416/

The Three Pillars of Provenance

1

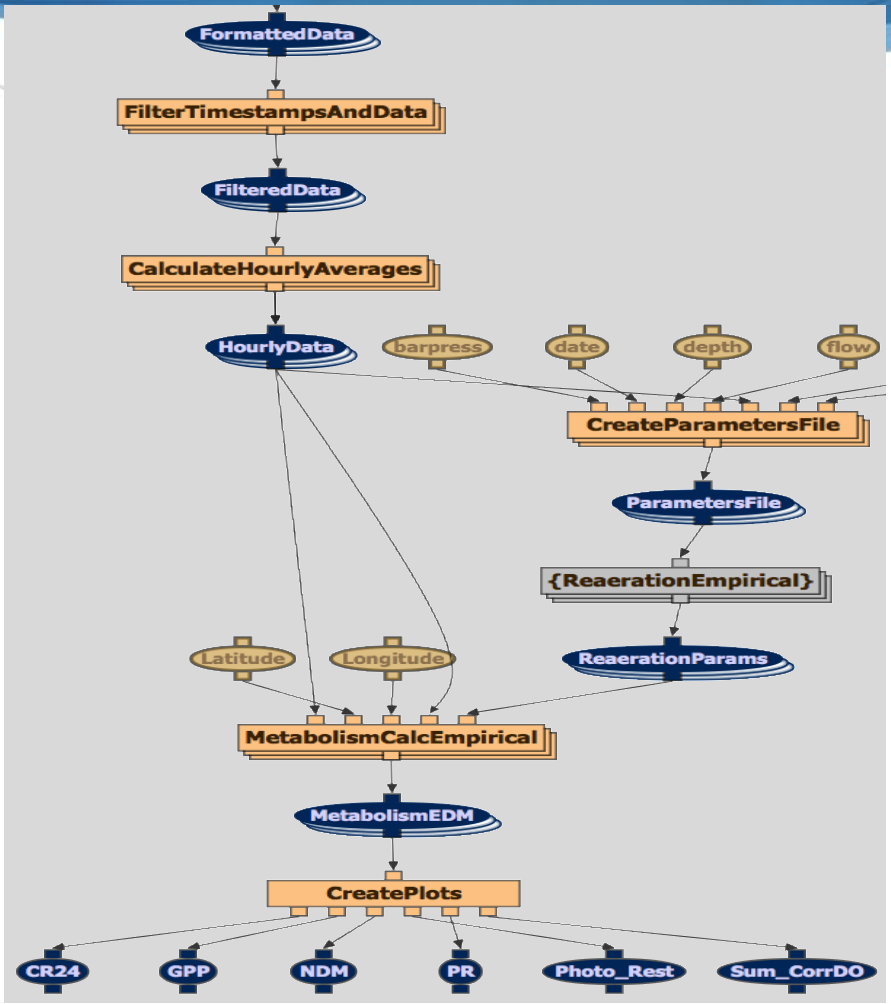
3

2

1. Processes
2. Resources
3. Attribution

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)
http://www.imdb.com/title/tt0040416/

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

Article Talk

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]

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.

References [edit]

1. ^ This article incorporates public domain material f Retrieved 2009-01-19.
2. ^ "Garibaldi volcanic belt: Garibaldi Lake volcanic field" 2010-06-27.
3. ^ Barlow, Nadine (2008). *Mars : an introduction to its int* ISBN 9780521852265.
4. ^ Stewart, Emily M.; Head, James W. (1 August 2001). *Geophysical Research* **106** (E8): 17505. doi:10.1029/20
5. ^ *a b c d e f g h i j k l m* This article incorporates public Jacquelyne; Tilling, Robert I. "Plate tectonics and people

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	CREUTERS journalist
Subject	
Description	
Publisher	FA Times
Contributor	Duckingham Palace
Date	
Type	
Format	
Identifier	
Source	original CREUTERS 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 trou
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

http://www.w3.org/2005/Incubator/prov/wiki/What_Is_Provenance

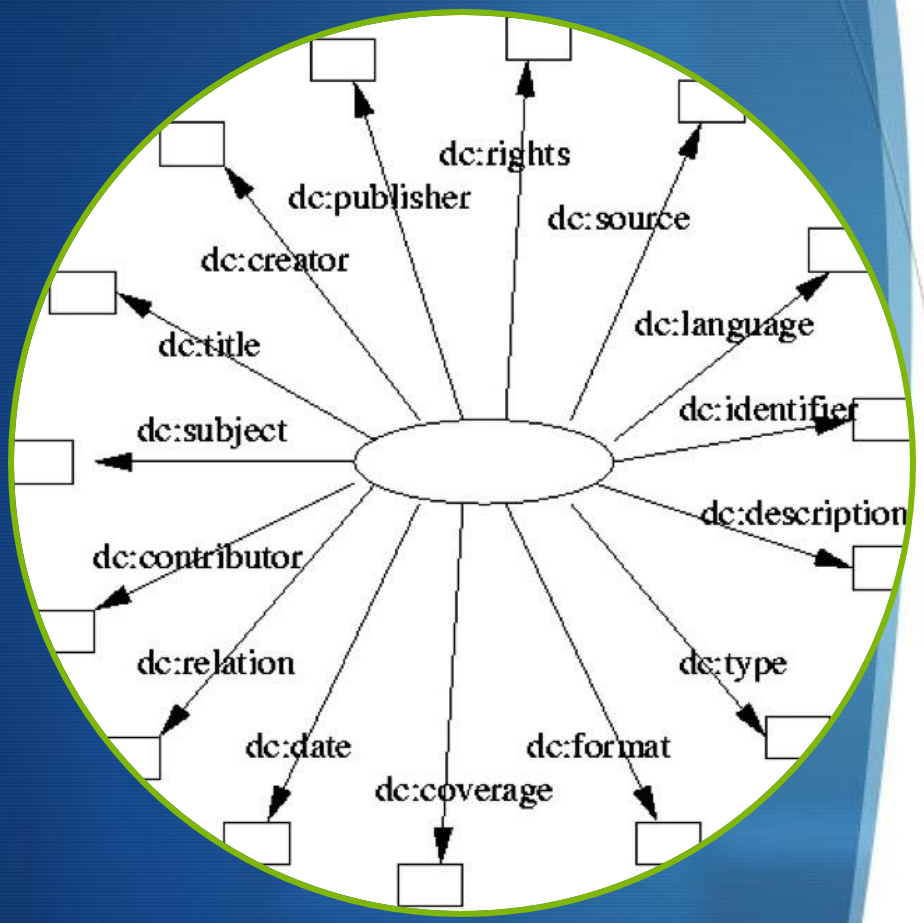
A Working Definition of Provenance

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



A Well-Known Provenance Vocabulary:

The Dublin Core

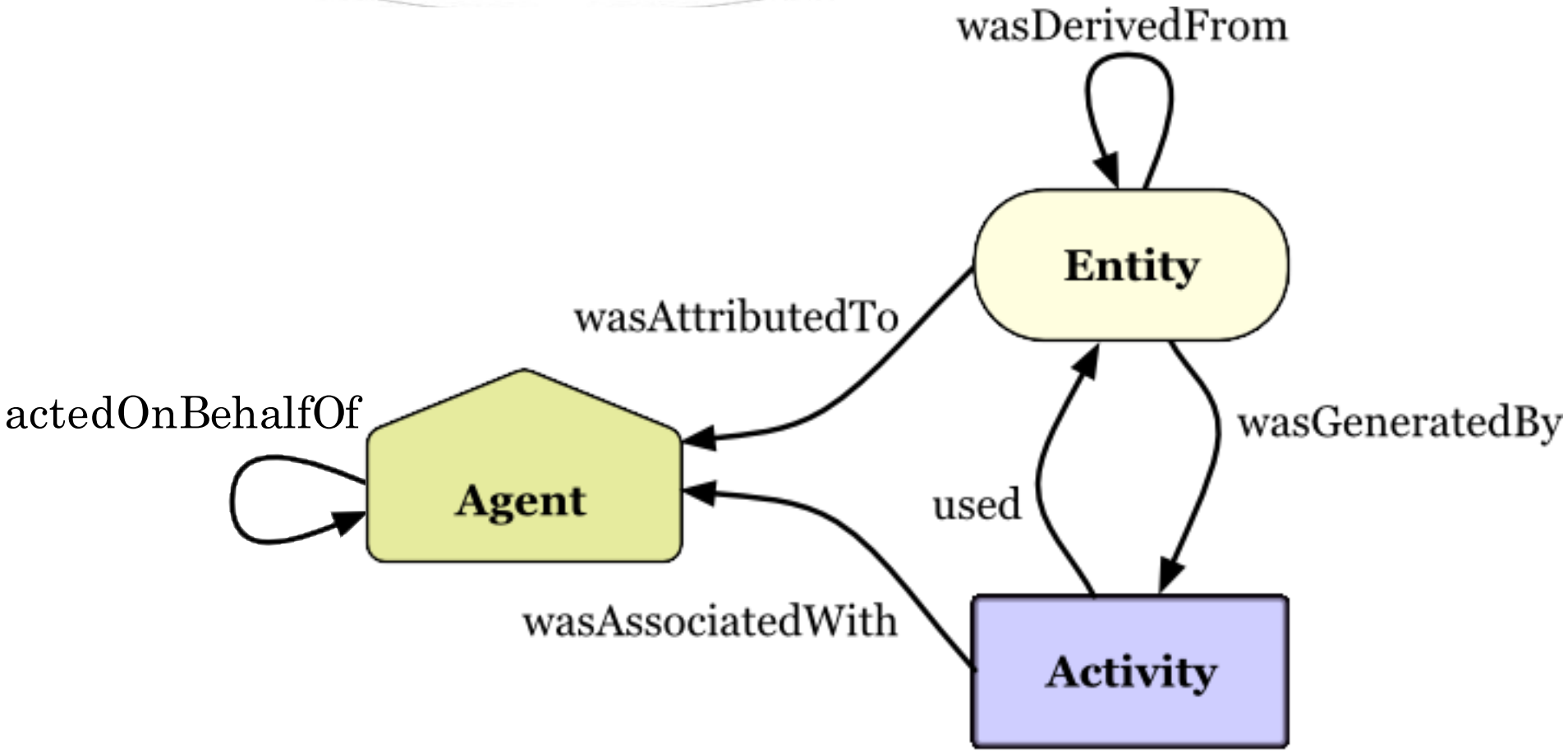
From library sciences

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

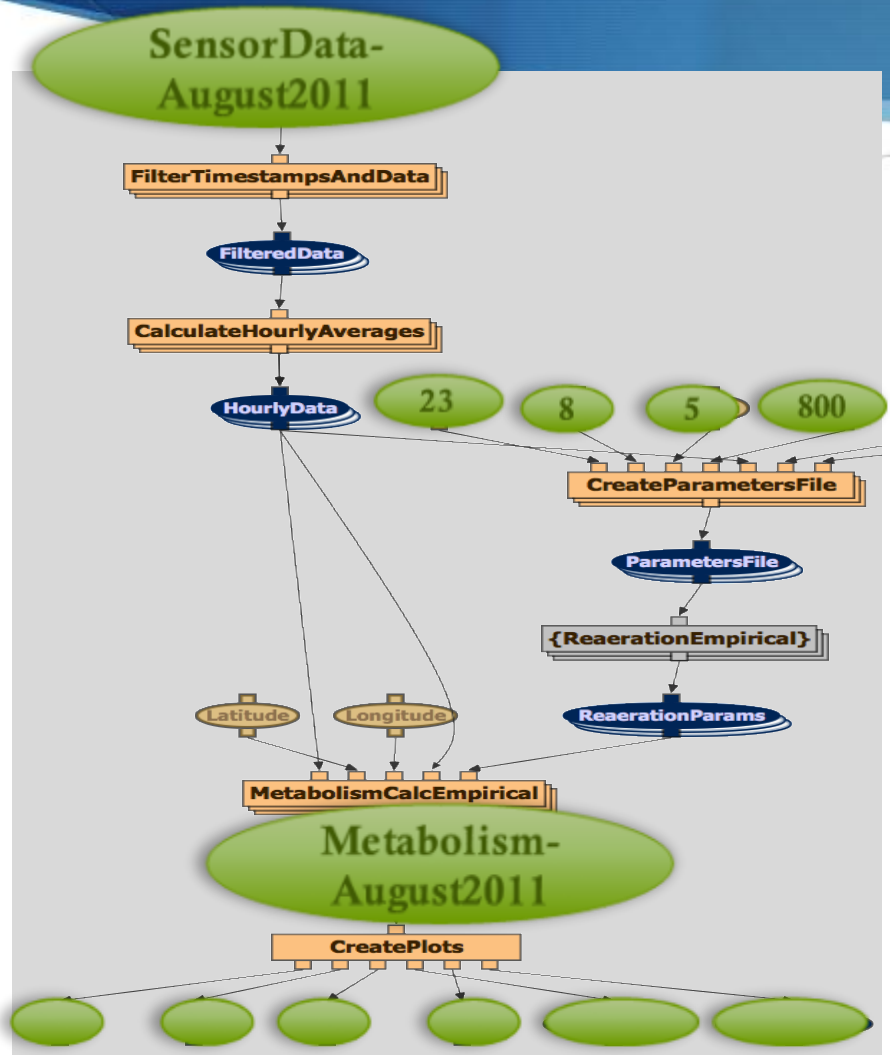
http://dublincore.org/documents/dcq-rdf-xml/

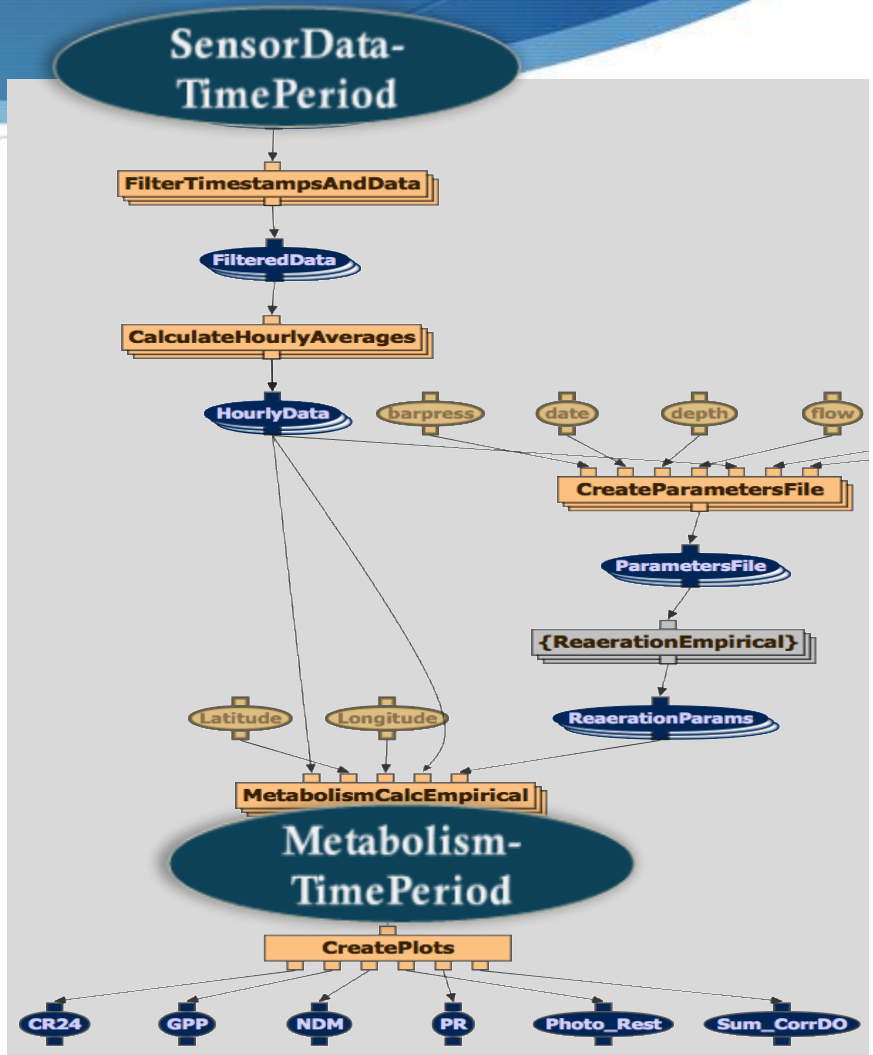
http://www.w3.org/TR/prov-primer/

A Provenance Standard for the Web: W3C PROV

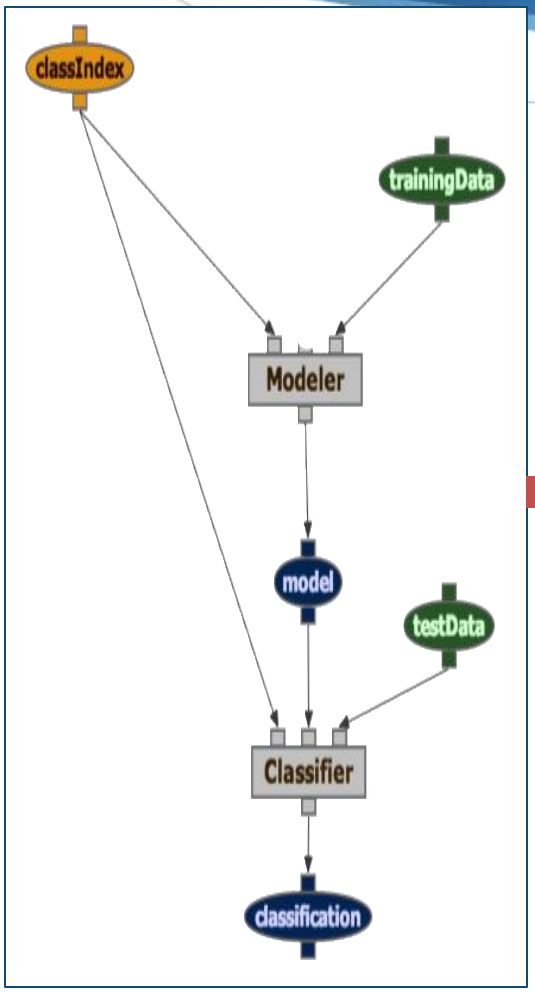


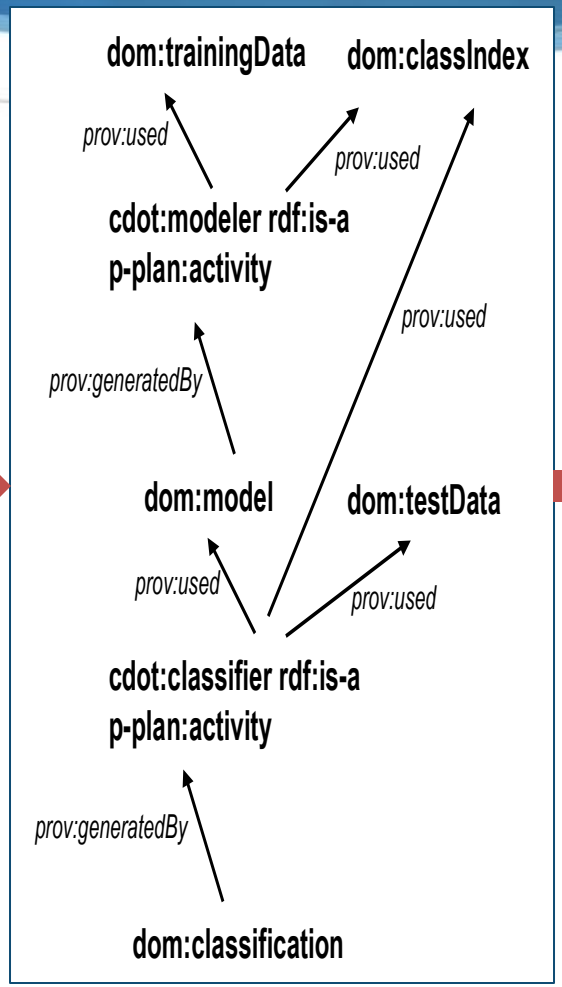
Describing Execution (Provenance) vs General Method (Workflow)





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 .
```

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

myexperiment

crowdLabs

SHIWA Workflow Repository
Click to visit

OPMW Workflow repository

An Example

Understanding kinematic data from the Hellerman thrust zone

Jade Silverstein

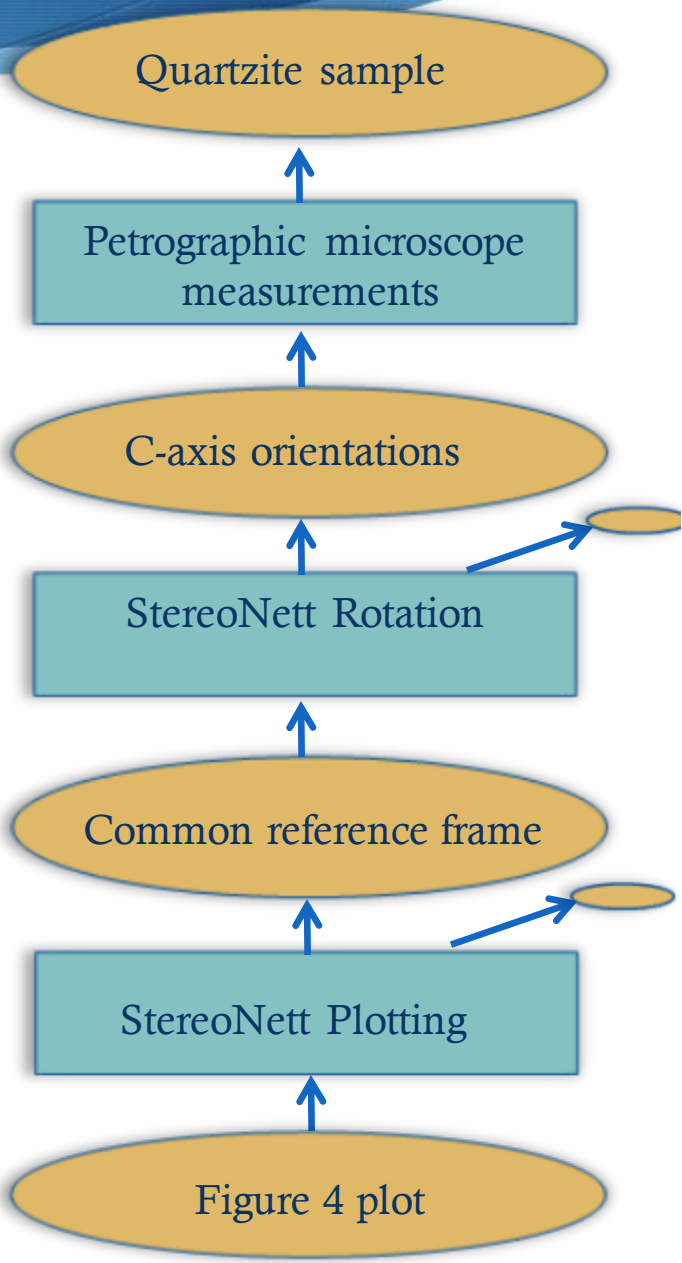
[…] We took a quartzite sample from the Hellerman thrust zone, and cut 3 thin sections. We measured c-axis orientations using a petrographic microscope. We rotated to a common reference frame using Duyster's StereoNett program. We plotted the data on lower hemisphere, equal area projections using Duyster's StereoNett program, shown in Figure 4. […]

An Example: Workflow

Understanding kinematic data from the Hellerman thrust zone

Jade Silverstein

[…] We took a quartzite sample from the Hellerman thrust zone, and cut 3 thin sections. We measured c-axis orientations using a petrographic microscope. We rotated to a common reference frame using Duyster's StereoNett program. We plotted the data on lower hemisphere, equal area projections using Duyster's StereoNett program, shown in Figure 4. […]

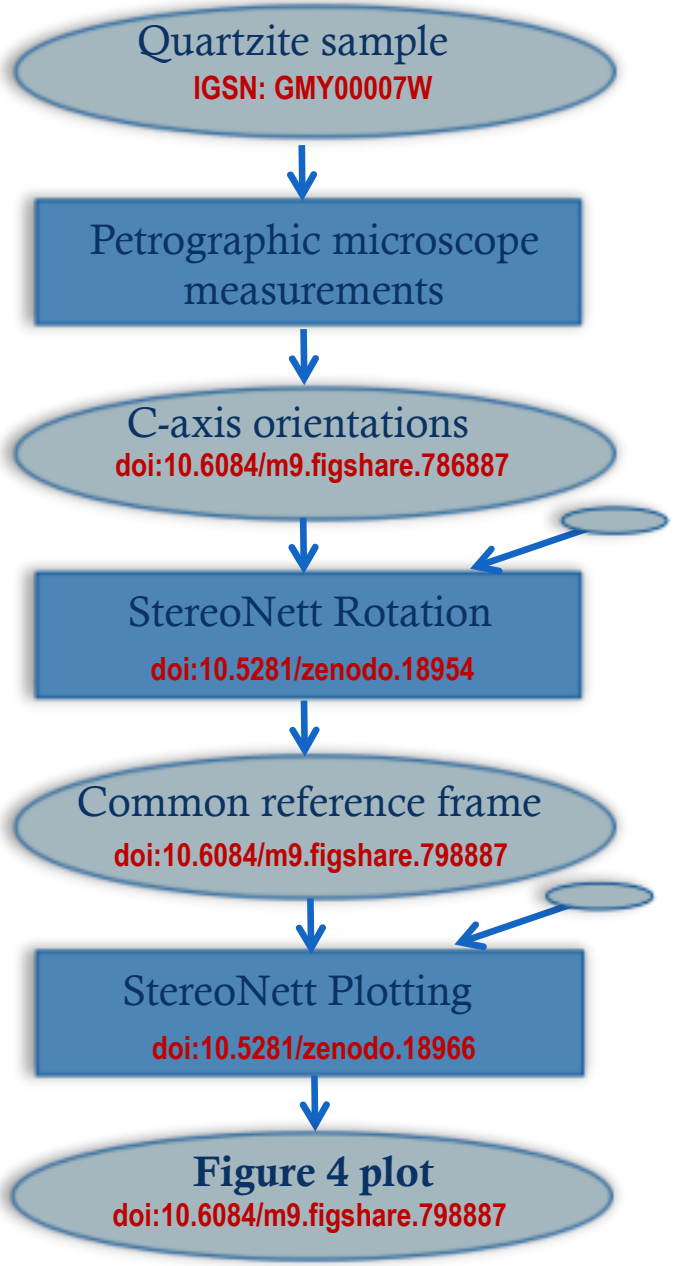


An Example: Provenance

Understanding kinematic data from the Heller 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 Heller thrust zone, 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. […]



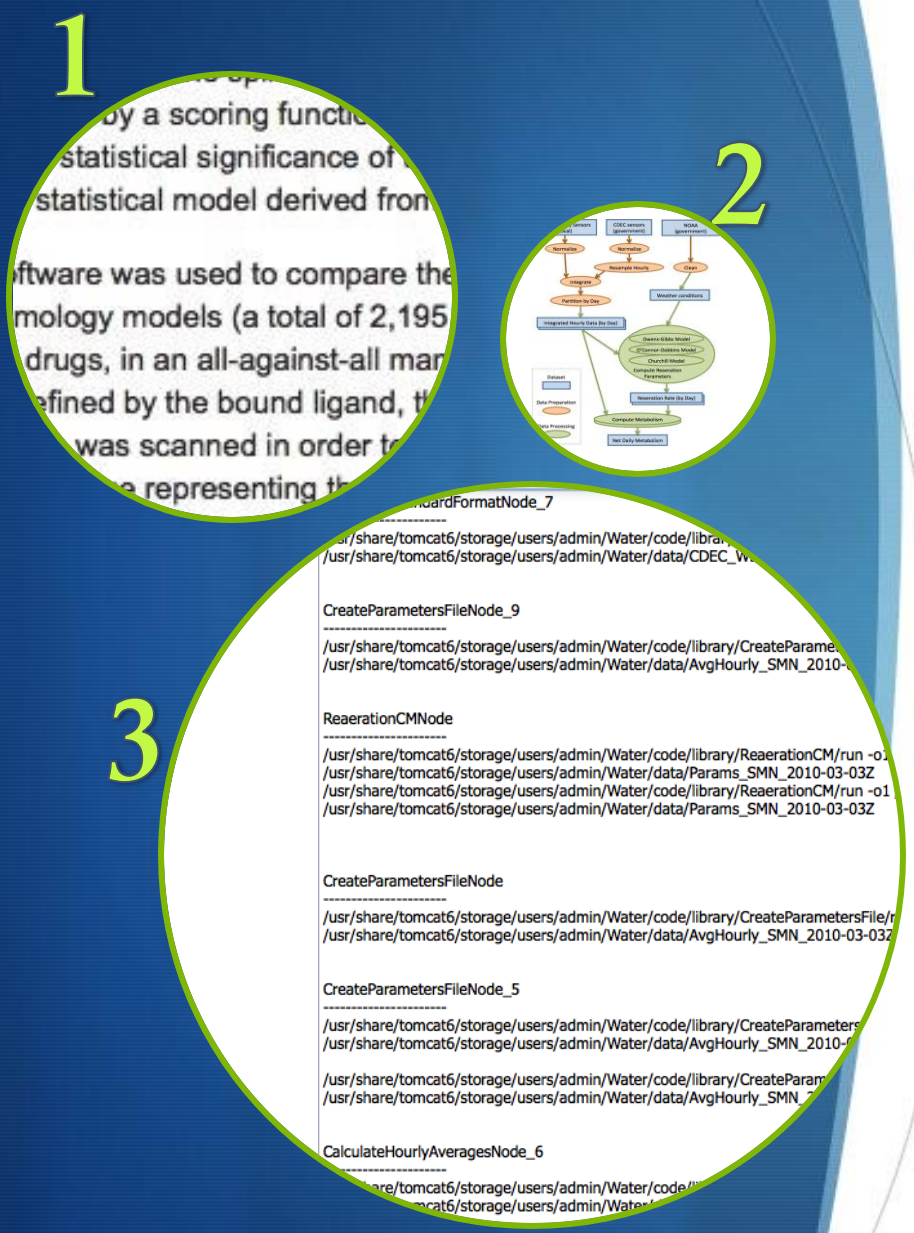
Goals of this Section

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

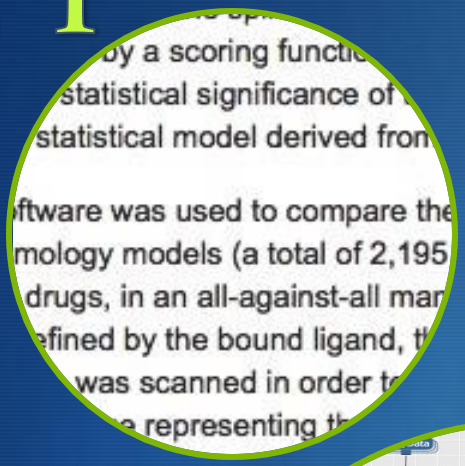
1. Describe the workflow in text
   - Data + software + workflow
   - Specify unique identifiers for data and software, versions, credit all sources
2. Develop a workflow sketch
   - Capture high-level dataflow across components
3. For provenance, include a summary or an execution trace

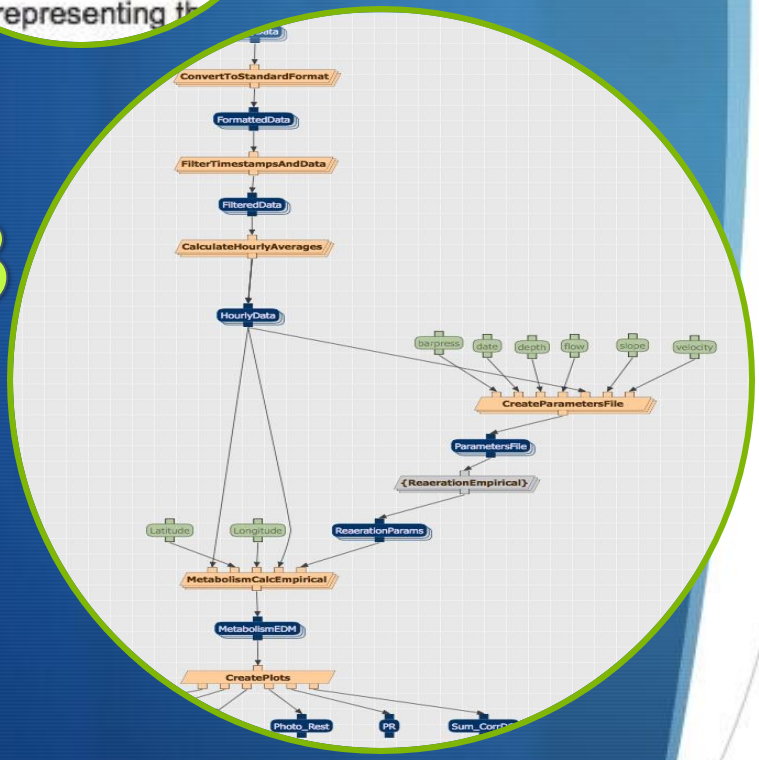


Documenting Provenance and Methods:

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 (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 an 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)

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. *PRACTICAL EXERCISE*
5. Documenting software with metadata

Part II

1. Documenting software dependencies
2. Documenting methods and workflows
3. Documenting provenance
4. *PRACTICAL EXERCISE*
5. Summary of author checklist

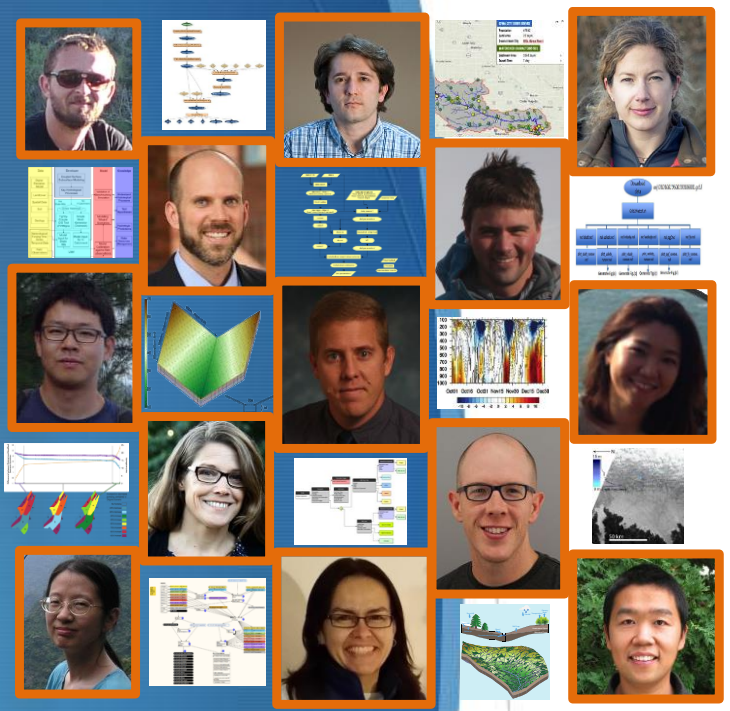
PRACTICAL EXERCISE: Representing Provenance

- Laura designs a survey about student financial support
- Jack and Jill conduct the survey and collect data
- A year later, Laura revises the survey
- Peter and Paula conduct the survey and collect data
- Zack compiles all the survey results, analyzes them with a statistics package, and publishes a paper with Laura and other co-authors

Sketch a diagram using PROV for:

1. Entities
2. Activities
3. Use and generation
4. Agents
5. Revision and derivation
6. Plans

The Scientific Paper of the Future: An Author Checklist



Part 2.5

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

http://www.scientificpaperofthefuture.org

ICER-1440323
ICER-1343800

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

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 Research

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)

Author Checklist

1. Data accessibility
2. Data documentation
3. Software accessibility
4. Software documentation
5. Provenance documentation
6. Methods documentation
7. Authors identification

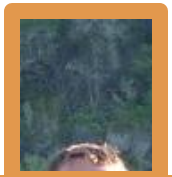
- **For datasets**, the paper should include one or more citations, specifying the authors, the site where they are described and can be accessed, the repository, and the license.
- **For software**, the paper should include one or more citations, specifying the authors, the site where it is described 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)

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), Daniel Garijo (UPM), 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), Gail Clement (CalTech), and Greg Wilson (Software Carpentry) for feedback on best practices
- Thank you also to the many people that have taken the training and asked hard questions
- We are grateful for the support of the National Science Foundation and the EarthCube program

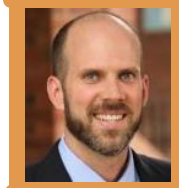
GPF Pioneer Authors



Cedric David, NASA/JPL
Hydrology modeling

Ibrahim Demir, U. of Iowa
Hydrology sensor networks

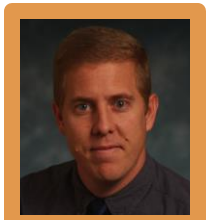
R. W. Fulweiler, Boston U.
Biogeochemistry in marine ecology



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

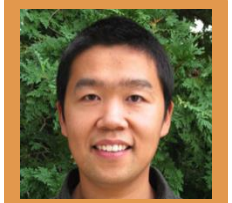
Ji-Hyun Oh, USC
Tropical meteorology

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

Published Articles

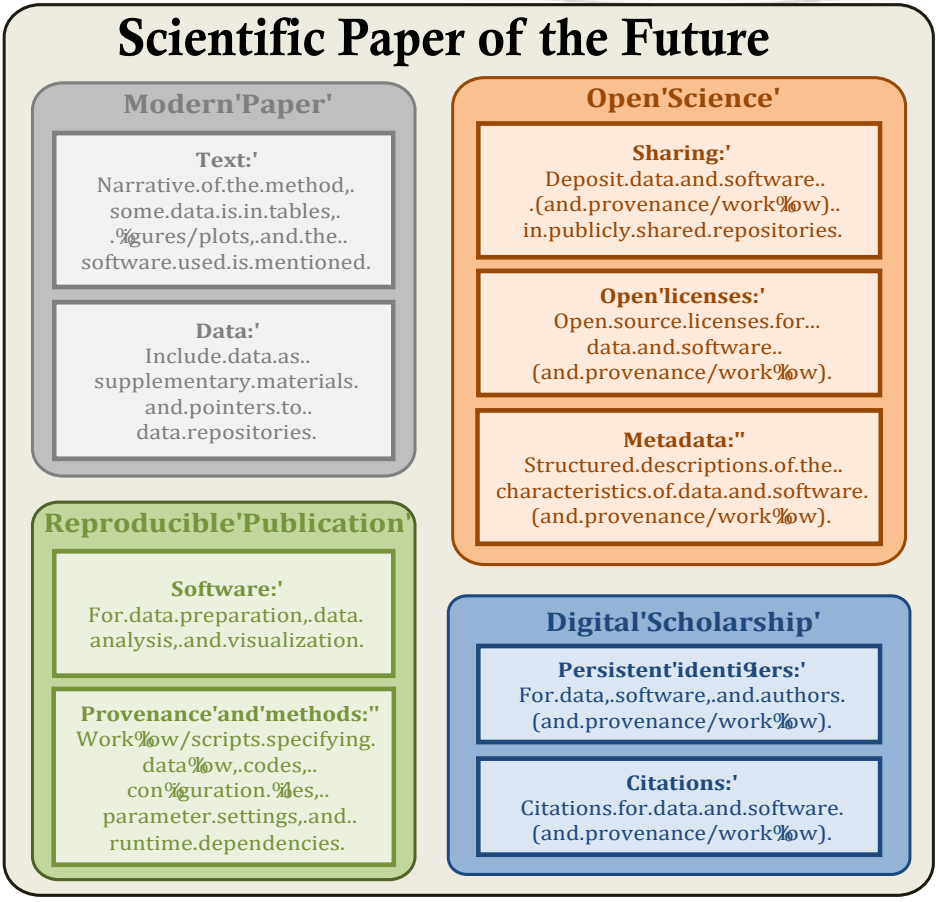
www.scientificpaperofthefuture.org/gpf/special-issue

"Towards the Geoscience Paper of the Future: Best Practices for Documenting and Sharing Research from Data to Software to Provenance" Gil et al, Earth and Space Science, 2016. http://dx.doi.org/10.1002/2015EA000136

- [David et al 2015]: 10 years of hydrology model software
- [Yu et al 2015]: Model coupling for surface/subsurface flow
- [Essawy et al 2015]: Hydrology workflows for reproducibility
- [Pope et al 2015]: Estimate subglaciar lake depth from imagery
- [Fulweiler et al 2016]: Long-term estuary data & products
- [Tzeng et al 2016]: Data processing for ocean observatory
- [Demir et al 2017]: Sensor network for flood monitoring
- [Peckham et al 2017]: Hydrological modeling toolkit

For More Information

http://www.scientificpaperofthefuture.org



Recommended best practices:
http://dx.doi.org/10.1002/2015EA000136

Special issue:
http://tinyurl.com/ess-gpf

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

ICER-1440323
ICER-1343800