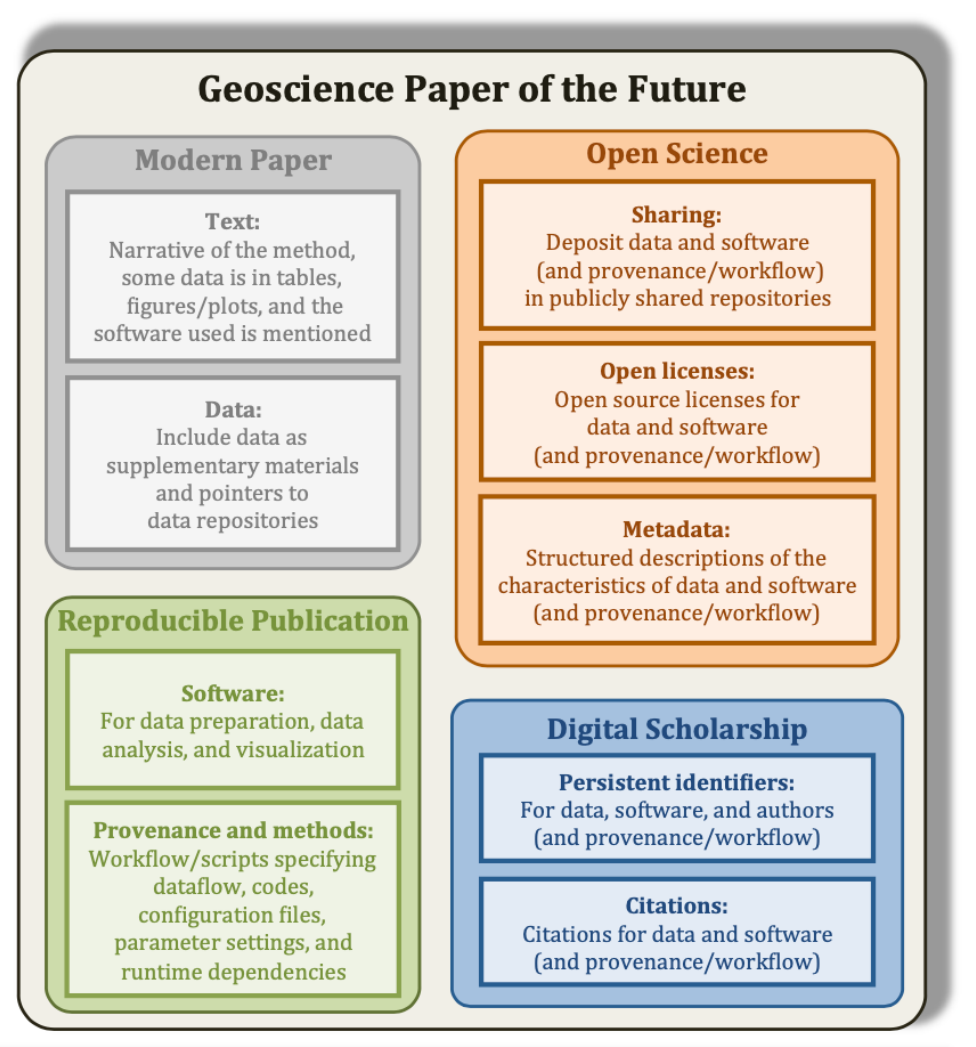
IS-GEO and iHARP 2022 Virtual Workshop



Documenting Software through Metadata

Presented by Suzanne Pierce

Slides by Yolanda Gil
USC/ISI
gil@isi.edu

May 1, 2022

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

Software Matters

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;

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

Software Matters

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

Software Matters

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 Matters

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.

Software Catalogs vs Software Registries

- **Software catalogs**
  - They are code repositories
  - Support software evolution
  - Support open source software development
- **Software registries**
  - They capture metadata
    - Useful structured information about your code

github
SOCIAL CODE HOSTING

CSDMS
COMMUNITY SURFACE DYNAMICS MODELING SYSTEM

figshare

Goals Today

1. Look at the Ontosoft example of **how researchers can share knowledge more easily by creating registries.**
2. Understand what metadata needs to be documented about software to promote reuse
3. Understand how to specify metadata to document software

Software Metadata

- Describe characteristics of the software that can be used to understand, discover (find), and compare software
- We will use OntoSoft as a model
  - Developed as part of the GeoSoft project
  - http://www.ontosoft.org

OntoSoft: An Ontology for Software Metadata

Identify

Locate – unique identifier
- has name (desc)
- has short description (desc)
- has software category (desc)
- has unique ID (uniqueID)
- has project web site (location+)

Understand

Relate – domain knowledge
- has domain keywords (desc)
- has uses and assumptions (desc)
- has use limitations (desc)
- similar software (desc)

Trust – quality and ratings
- has creator (agent+)
- has publisher (agent+)
- has major contributor (agent+)
- commitment of support (desc)
- has adopters (entity+)
- has use information (desc)
- has use statistics (desc)
- used in publication (citation+)
- has benchmark information (desc)
- has salient qualities (desc)
- has funding sources (desc)
- has rating (rating+)

Do Research

Experiment – run with other data
- has input (i-o)
- has input parameter (i-o)
- has output (i-o)
- has relevant data sources (desc)

Compose – run with other software
- has interoperable software (desc+)
- has composition description (composition)

Cite – scientific publications
- has preferred citation (citation+)

Execute

Access – download
- has code location (location)
- has executable location (location)
- has license (license+)

Install – execution requirements
- has documentation (location)
- has installation instructions (desc)
- has implementation language (language+)
- has dependency (software version)
- requires average memory (measurement)
- supports operating system (os)
- has average run time (desc)
- has other implementation details (desc)

Run – testing execution
- has test data (desc)
- has test instructions (desc)

Get Support

Discuss – support and community
- has email contact (email)
- has software support (desc)

Update

Track – evolution
- has software version (version)
- has version release date (date)
- supersedes (version)
- superseded by (version)

Contribute – evolution
- has active development (desc)
- has software community (desc)

1) Identifying Software

- Name
- Short description
- Unique identifier
  - (Permanent) URL
  - DOI
- General categories/keywords/labels/classes
- Project web site

github SOCIAL CODE HOSTING

sourceforge

CSDMS COMMUNITY SURFACE DYNAMICS MODELING SYSTEM

ESMF

CIG COMPUTATIONAL INFRASTRUCTURE for GEODYNAMICS

OSGeo Your Open Source Compass

Apache Open Climate Workbench

figshare

Main Types of Unique Identifiers

1. URL
2. Permanent 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/

Persistent URL (PURL)

- The same PURL can be resolved to different Web address over time
  - You always refer to your software with the same PURL: http://purl.org/mysoftwandme/awesomesoftw.html
    - Today you are in grad school and tell purl.org to resolve it to: http://www.wisc.edu/myadvisorsgroup/awesomesoftw.html
    - Tomorrow you have graduated and tell purl.org to resolve it to: http://www.stanford.edu/myowngroup/awesomesoftw.html
- It is easy to create your own PURLs, just remember to update whenever you move the software
  - Go to http://www.purl.org (and others)

PLoS Biol. 2003 Nov; 1(2): e57.
Published online 2003 Nov 17. doi: 10.1371/journal.pbio.0000057
PMCID: PMC261894

The What and Whys of DOIs

Susanne DeRisi, Rebecca Kennison, and Nick Twyman

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).

2) Understanding Software: Domain Knowledge

- Links to similar software
- Recommended uses and assumptions
- Constraints on its use with data, or other limitations
- Domain-specific keywords

2) Understanding Software: Trust

- Creator
- Major contributors
- Publisher
- Funding sources

- Use statistics
- Ratings
- Benchmark information

- Adopters
- Publications that used it
- Projects that adopted it

- Commitment of support

3) Execute: Access

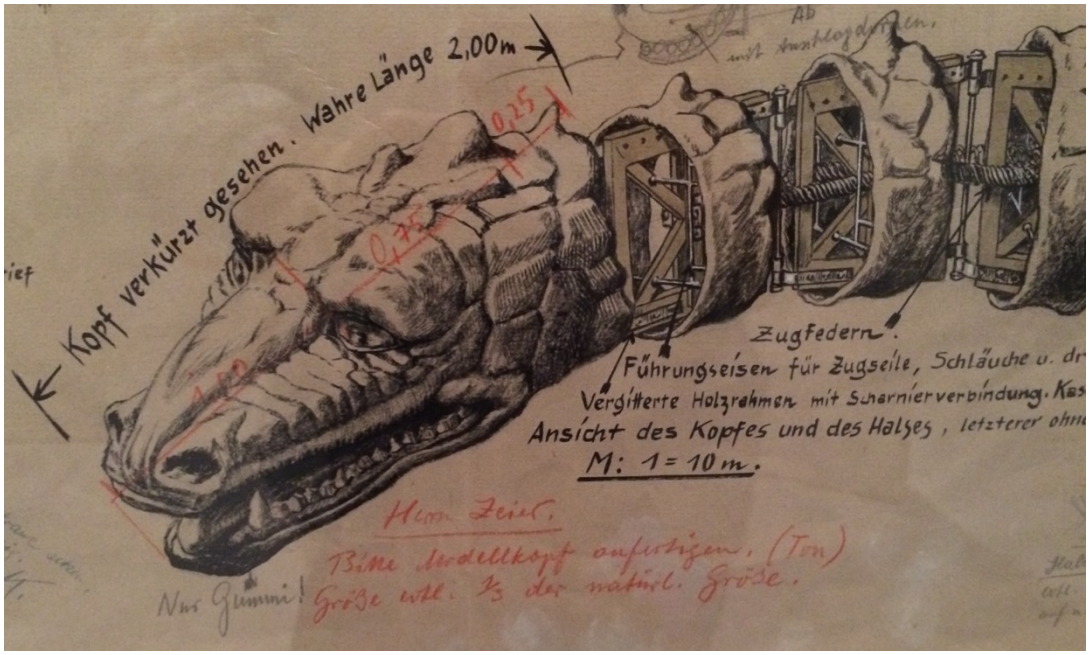
- License
- Code location
- Executable location

3) Execute: Install

- Documentation
- Installation Instructions
- Implementation language
- Dependencies on other software
- Memory requirements
- OS requirements
- Average run time
- Lots more implementation details possible: e.g., parallel implementations

3) Execute: Run

- Testing instructions
- Test data



4) Get Support

- Email contact to send questions or report bugs
- Software support provided

5) Do Research: Experiment

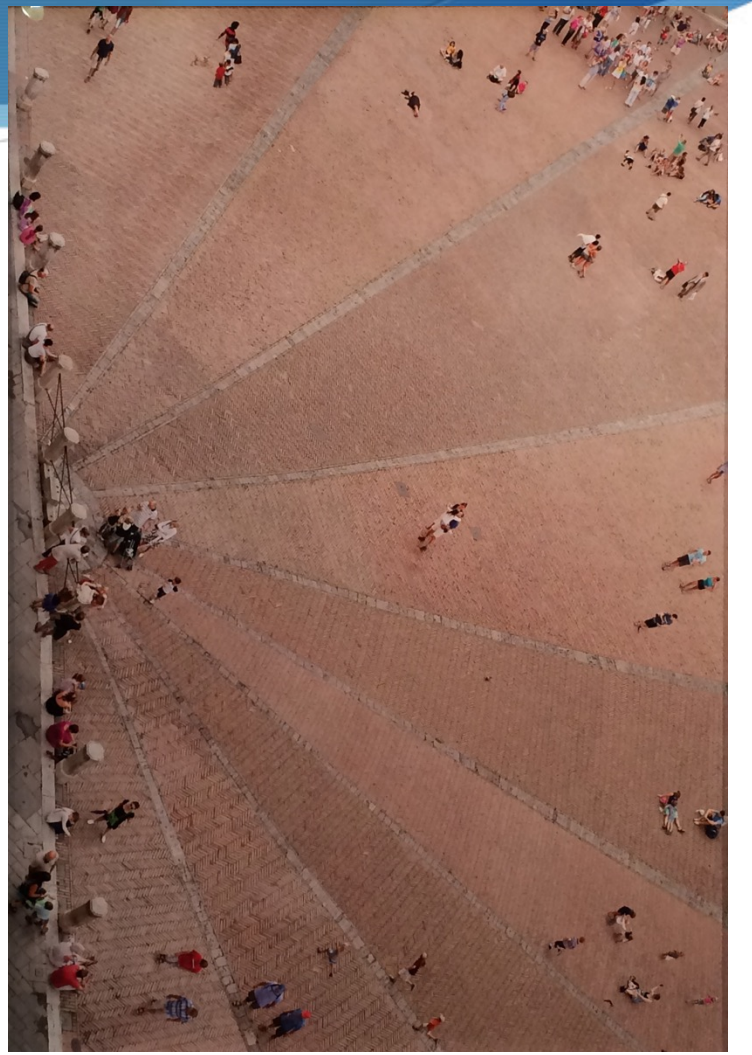
- Input data
- Parameters
- Output data
- Relevant data catalogs or data sources typically used with the software

5) Do Research: Compose

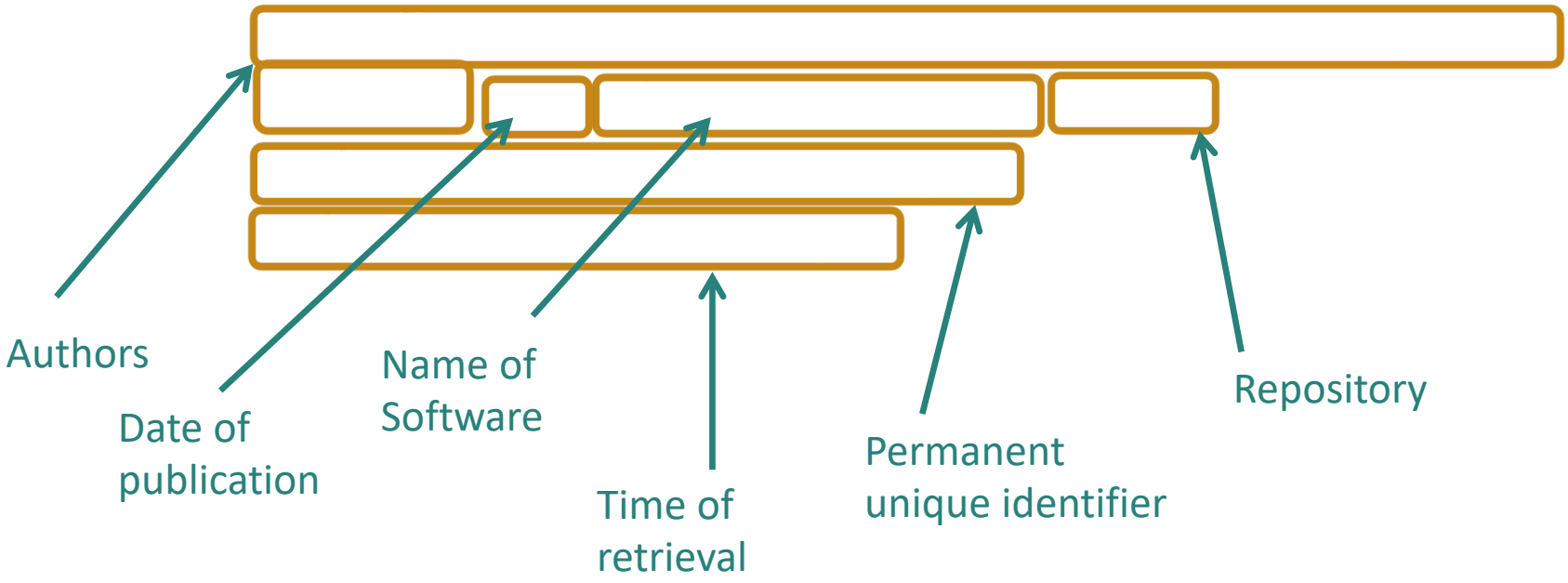
- Interoperable software
  - Data preparation software
  - Visualization software
  - Data post-processing
- Software composition
  - Workflow description

5) Do Research: Cite

- Preferred citation, could be based on:
  - A citation as a project
  - A publication
  - A DOI



Software Citation



Software Papers

- Some journal articles describe a piece of software
- Some publications have "software papers"
- See also Journal of Open Research Software: "software metapapers"

6) Update: Track

- Software version
- Version release date
- Supersedes
- Superseded by

6) Update: Contribute

- Active development
- Planned releases
- Software community

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 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
- … the software is set up to run in a local cluster or cloud?
  - Provide instructions for deploying it elsewhere

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…"

How to use the software citation in your paper?

- 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
  - With an in-text pointer in a special “Software Resources” section
  - With an in-text pointer in the Acknowledgements section

Goals Today

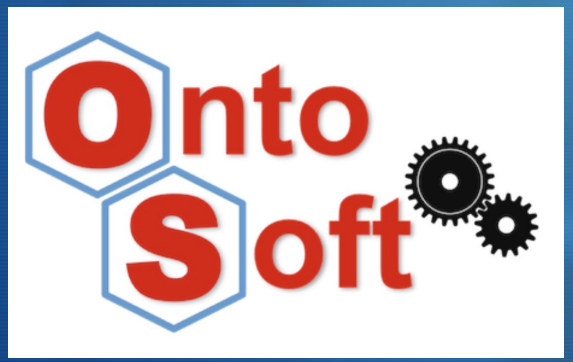
1. Understand what needs to be documented about software to promote reuse
2. Understand how to specify metadata to document software

Suggested Approach

1

2

3



1. Create a public entry for your software with a permanent unique identifier
   - Go to github.com, create an account
   - Create an entry for your software with README, etc.
   - Specify a license -- choose from http://opensource.org/licenses
2. Use OntoSoft to create a detailed software description in the IS-GEO Registry
   - https://www.ontosoft.org/