Geoscience Paper of the Future: Learning Best Practices for Scholarly Publication **[1]**

Submitted by annieburgess on Fri, 2015-06-19 15:35 Thursday, July 16, 2015 - 13:30 to 15:00
Thursday, July 16, 2015 - 15:30 to 17:00
Event: Summer Meeting 2015 [2]
Session Type: Breakout [3]
Expertise Level: Beginner [4]
Abstract/Agenda:
The Geoscience Paper of the Future (GPF) is an initiative to encourage geoscientists to publish papers together with the associated digital products of their research. This means that a paper would include: 1) Documentation of datasets, including descriptions, unique identifiers and availability in public repositories; 2) Documentation of software, including pre-processing of data and visualization steps, described with metadata and with unique identifiers and pointers to public code repositories; 3) Documentation of the provenance and workflow for each result.

The GPF Initiative has two major components:

1. A Special Issue of the AGU Earth and Space Science journal that will highlight GPFs. We have partnered with the largest scientific society in geosciences to create a special issue of their new cross-disciplinary journal to encourage submissions of GPFs that will help move the geosciences community to publish all digital objects resulting from their research. The submission deadline is January 1, 2016.

2. The GeoSoftCamp program offers in-person and online training sessions for geoscientists to learn best practices in software and data sharing, provenance documentation, and scholarly publication. The training sessions will be offered at community events throughout 2015. In addition, the ESIP Federation will support the extension of the training sessions into independent learning resources such as video tutorials, podcasts and other formats, and made available in 2015 and 2016.

The training is divided into two sessions, 90mins each. Training materials are provided to all participants.

For each training topics, basic concepts and best practices are explained. A summary at the end provides specific advice and pointers to implement those best practices.

The GPF Initiative was initiated by the EarthCube GeoSoft project with funding from the National Science Foundation.

More details about the GPF Initiative are available at http://www.geosoft-earthcube.org/gpf/ [5].

Geoscience Papers of the Future I: Making data and software accessible

- What are the benefits of augmenting papers with data, software and provenance that are properly documented and cited
- How to publish data in a public shared repository, use a license and cite it in an article
- How to publish software in a public repository, use a license and cite it in an article

Geoscience Papers of the Future II: Describing software and describing provenance

- How to publish data in a public shared repository, use a license, and cite it in an article
- What are the benefits of augmenting papers with data, software, and provenance that are properly documented and cite
- How to publish software in a public repository, use a license, and cite it in an article

Notes:
Part 1

- The presentation is available at the following link: http://www.geosoft-earthcube.org/ [6]
- The Geoscience Paper of the Future: OntoSoft GPF Training:
    - There will be a special issue available from AGU.
        - For further information:
            - http://tinyurl.com/ess-gpf [7]
            - http://www.ontosoft.org/gpf [8]
    - Important to integrate the workflow as natural as possible as part of the researchers' "typical activities".
- Motivation and Overview:
    - Traditional articles when published might include data, but rarely include associated software.
    - The more open science is, the more likely new discoveries and contributions can be made from additional sources.
    - Scientists place heavy emphasis on development of research resources for the scientific community; however, the importance of data deposit is still not being recognized.
    - Change in policies is slowly helping to encourage the scientists to share their data.
        - However, there are still a lot of unpublished data (dark data).
        - Currently, publications of data are being encouraged, but software is still only loosely documented.
    - There are many "excuses" that researchers have for not sharing software.
        - This caused software to become "dark" as well.

- Reproducibility affects the following areas: methods, reliability, scientific integrity, trust, financial, human lives.
- Other issues that are important relating to software publication: licenses, citation, and
- Three key areas of Geoscience Papers of Future (GPF): data, software, and provenance.
- The GeoSoft GPF Pilot was conducted and 13 papers were created initially.
- New training sessions will also be available after the ESIP Summering Meeting.

- Making Data Accessible:
  - We are getting better at producing data paper and depositing data, but data sharing can still be improved.
  - Best practices:
    - Publication in a repositories.
      - Options between not curated vs. curated repositories.
      - Directories of research data repositories are also available.
    - General and domain metadata.
      - General vs. domain specific.
      - License is a specific element that is important when sharing license
        - One example is Creative Commons.
    - Accessibility of Data
      - Manual vs Machine.
    - Permanent and unique Identifiers:
      - URL/URI - persistence is key trade-off.

- PURL - Persistent URL; update is still important.
- DOI - Figshare and GitHub are examples of organizations that could provide DOIs.

- Citation preference:
- Make it easy for people to cite by providing the citation.

- It is important to emphasize that obtaining a DOI is a professional commitment. As a result, it is really important to plan for the sustainability of the dataset.
- AGU's data management maturity is focusing on sustainability.
- Based on CMMI Institute's DMM Structure.
- Measures both capability and maturity.
- Full roll out of the program will be available in fall, 2015.

- Making Software Accessible:
- There are 5 recommended best practices:
- Making software executable by others:
- Portability = how many different environments can the software be run?
- This included all the required dependencies.
- Modularize the code, so that the software can be run using a workflow.
- Also, allowing software to be configurable and testable can help the software to be more understandable (this includes providing understandable error messages).

- Preparing source executable:
- Provide the source code can help improve transparency.

- Alternatively, executable files can be run more quickly.
    - Important to include all the dependencies and document files (e.g. ReadMe).

- Open source software publication:
    - Defining licence type is again very important just like data publication.
        - Open source initiative: http://opensource.org/licenses [9]
    - The apache Software Foundation is an example of software publication site.

- Permanent and unique identifier:
    - Similar options as for data.

- Citation preference
    - Including the version information is recommended.
    - The recommended format is based on ESIP's data citation format.

Part 2

- Documenting software with metadata:
    - Software repository is not the same as software registry.
        - Software registry is meant to capture metadata while software repository focuses providing the actual code.
        - It is important that the software is registered as well as deposited.
    - OntoSoft has six major categories of software metadata (http://www.ontosoft.org [10]):

- Identification of software:
    - Name, Short description, unique identifier, keywords, and web site.
- Add domain knowledge software:
    - Links to similar software, recommended uses and assumptions, constraints/limits, domain specific keywords.
- Build software trust:
    - Ex: Creator, major contributors, publishers, publications that used the software, commitment of support, use statistics, adopters.
- Execution:
    - Access of the software for execution:
        - License, code location.
    - Installation of the software for execution:
        - Ex: average run time, OS requirement, memory requirement, and additional implementation details.
    - Actually running the software:
        - Testing instructions and test data.
    - How to get support:
        - Contact information.
- Do Research:
    - When doing experiment with the software:
        - Ex: Input data, parameters, output data.
    - Compose:

- Interoperability and software composition.

- Cite:

- Provide preferred citation style.

- Update:

- Track:

- Ex: Software version, version release.

- Contribute:

- Ex: planned releases, software community (such as mailing list).

- Additional helpful information that could help the software to be discovered.

- Provenance and workflow:

- Provenance is science, and it is becoming increasing more important because of the focus on reproducibility.

- Workflows need to be first class citizens in science.

- Crucial in enabling and upholding the reproducibility of science.

- It is important to note any potential issues in the workflow, such as services provided by third parties.

- Publication of workflow will have significant positive impact on transparency.

- Process, document and entities are three key manifestation of provenance.

- Provenance can be seen as metadata, but not all metadata is provenance.

- W3C Prov: a provenance standard for the web.

- Steps for executing a software is not necessarily the same as the steps for general

methods for getting the software to work.

- Suggested approach:
  - Step 1: Describe the workflow in text
  - Step 2: Develop a workflow sketch
  - Step 3: Specify the formal workflow
- Workflow should also receive a DOI and be included in a separate “Methods” section.
  - It can be published as a separate document but linked to the software.

- Summary of GPF author checklist:
  - Elements of the checklist:
    - Data Accessibility
    - Data Documentation
    - Software Accessibility
    - Software Documentation
    - Provenance Documentation
    - Methods Documentation
    - Author Identification
- 2 use cases were presented.

Session Leads:

Name: Yolanda Gil [11]
Organization(s): USC/ISI [12]

Presenters:

Name: Yolanda Gil [11]
Organization(s): USC/ISI [12]

Notes takers:

Name: Sophie Hou [13]
Organization(s): UCAR/NCAR [14]

Participants:
Part 1 -

In Person: Yolanda Gil, Ji-Hyun Oh, Sandra Villamizar, Shelley Stall, Ruth Duerr, Stefan Falke, Nancy Hoebelheinrich, Audrey Mickle, Grace Peng, Brian Wee, Steve Aulenbach, Victor Zlotnicki, Karl Benedict, Nikunj Oza, Nic Weber, Sophie Hou

Remote: Kent

Part 2 -

Yolanda Gil, Ji-Hyun Oh, Sandra Villamizar, Shelley Stall, Grace Peng, Victor Zlotnicki, Shannon Rauch, Danie Kinkade, Ge Peng, Nancy Hoebelheinrich, Audrey Mickle, Stace Beaulieu, Karl Benedict, Nick Weber, Sophie Hou, Sky Bristol

Creative Common License: Creative Commons Attribution 3.0 License
Accepted:

Source URL: https://commons.esipfed.org/node/8075

Links
[1] https://commons.esipfed.org/node/8075
[2] https://commons.esipfed.org/2015SummerMeeting
[3] https://commons.esipfed.org/session-type/breakout
[4] https://commons.esipfed.org/taxonomy/term/260
[5] http://www.geosoft-earthcube.org/gpf/
[6] http://www.geosoft-earthcube.org/
[7] http://tinyurl.com/ess-gpf
[8] http://www.ontosoft.org/gpf
[9] http://opensource.org/licenses
[10] http://www.ontosoft.org
[11] https://commons.esipfed.org/node/484
[12] https://commons.esipfed.org/taxonomy/term/319
[13] https://commons.esipfed.org/node/7872
[14] https://commons.esipfed.org/taxonomy/term/2486