Citing a Black Box and its Output **[1]**

Submitted by jhausman on Wed, 2015-04-22 13:48 Tuesday, July 14, 2015 - 13:30 to 15:00
Event: Summer Meeting 2015 [2]
Session Type: Breakout [3]
Expertise Level: Beginner [4]
Collaboration Area: Information Quality [5]
Preservation and Stewardship [6]
Products and Services [7]
Science Software [8]
Abstract/Agenda:
Citing data is becoming more significant as science journals are requiring authors to provide their data sources. Data provenance is recognized as important for its stewardship and for reproducibility of experiments. But, what happens when you obtain your data through a tool or service that changes the data or metadata from the original file? This session will look at how to cite data that is produced from a tool that is essentially a black box to the user. This includes tools and services that subset, aggregate, calculate anomalies or climatologies, etc. Identifiers that can assist in making the citation more exact will also be investigated.

Agenda

Robert R. Downs - Why We Should Care About Citing Software and Subsetted Data

Shannon Rauch - Supporting Citation of Subsetted Data: A Data Center's Perspective

Ji-Hyun Oh and Yolanda Gil - Geoscience Paper of the Future Initiative

Jessica Hausman - Discussion on Citing Software and Subsetted Data

Session notes document: https://goo.gl/0ZkxEH [9]

Notes:

Citing a Black Box and its Output

Tuesday, July 14, 2015 - 13:30 to 15:00
Summer Meeting 2015
Chairs: Jessica Hausman, Bob Downs, Daine Wright

Bob Downs - Software and Data Subsets to Cite

- web services, data layers, subsets, map segments and Data Subsets
- Tools used for analyzing and subsetting data should be recognized for their contributions; standard software like Excel not necessary to cite.
- Software and data are disseminated as research products.
- Why cite software? Reproducibility, attribution, recognition.
- Different roles for citing software and subsetted data: data & software producers, data providers, data users, publication editors - each with different roles/responsibilities.
- Software and subsetted data should be referenced in Reference section just like other scholarly publications.

- web services, data layers, subsets, map segments
- tools that are specific to the work we are doing
    - citing generic 'office' type tools is not necessary
    - tools that contribute to science
- goal is to offer transparency and possibility of replication and comparison
- these subsets and tools are publicly disseminated products, possibly published
    - should be cited
- offer provenance for work we are doing
- boiled down to three general reasons: reproducibility, attribution, recognition
    - citing software including version to duplicate conditions of study
    - citing entire dataset does not give the whole story
    - attribution is the norm in science, and owed to tool and subset authors
    - recognition justifies effort of others and rewards their accomplishments
- roles: data and software producers, data providers, data users, pub editors
- Citing software and data makes them first class objects.
    - If we do not cite software and data used, how can we expect others to do so?
- Incentives for citing software & data subsets from publishers, government, software producers
    - what about hiring and tenure review committees
- related ESIP activities: data steward committee, science software cluster

Shannon Rauch - Supporting Citation of Subsetted Data: A Data Center's Perspective

- Slides: http://slides.com/shannonr/supporting-citation-of-subsetted-data-a-data-center-s-perspective [10]
- Data manager BCO-DMO, Woods Hole
- subsetted data is challenging to talk about
- Why cite data? Starr et al 2015, transparency, verification and reproducibility, attribution, discovery, re-use
- 8 core principles of data citation JDDCP, Force 11
- Minimum: Creator. (Year) Title. Provider. Identifier.
    - other fields highly recommended
- Citation is easier said than done, acknowledgement shows up in text or captions rather than references.
- Citation of full data set is often not even enough.
- Data are not usually static
    - dynamic data sets include versioning and subsetting, What about growing data sets?
- Researcher wants to but data set in version control box, so it can be referenced by version, but the data set stays together
- Deep citation, citation of subset, is analogous to citing page numbers in a book.
- Approaches to citing subsets
    - save each subset as unique data set: inefficient, redundant
    - cite entire data set, describe subset in paper: not good for machines
    - assign identifier to query
- what is level of granularity, consider machine needs
- RDA working group on Data Citation (WGDC) - two step approach
    - version kept in versioned, timestamped manner
    - assign persistent id (PID) to time-stamped queries
- BCO-DMO - 7500+ data sets, only most recent version available online
    - work on citations: Ocean Data Publication Cookbook
    - basic guidelines on website, but don't' include versioning and subsetting
- Subsetting example: data from three cruises
    - quick subsetting tool can cut data to two cruises

    - url appended with query, but not ideal
    - could assign PID to query
- Changing the culture is as important and putting technology in place.

Questions -

- Some journals are now requiring references to data sets. Is that being seen in your field?
    - Publications are now saying data must be explicitly cited.

Question/Comment from Ruth: AGU is now requiring citation of datasets and asking reviewers to look for those. Are other journals?

Ji-Hyun Oh - Geoscience Paper of the Future Initiative

- GeoSoft Project
- The value of software cannot be underestimated.
- Why is software not shared? - not confident in code, work for government, want to commercialize, etc.
- GeoSoft project is promoting best practices for software sharing.
- There is a software registry where anyone can register software and provide metadata about it.
- Geoscience Papers of the Future (GPF) is about data, software, and provenance. Data (metadata) and software are stored in public repository. Not only modeling software - but also other ancillary software for data formatting, conversions, etc.
- Call for papers for Special Issue of AGU's ESS Journal (call for papers opened July 2015).
- Permanent unique identifiers (doi) can be obtained for code through GitHub.
- ESIP Training Session Thursday 1:30 to 5:00 PM.
- http://geosoft-earthcube.org/gpf/ [11]
- http://onlinelibrary.wiley.com/journal/10.1002/%28ISSN%292333-5084/homepage/call_for_papers.htm [12]

- Practical aspect of software citation
- GeoSoft: Software Stwaredship for Geosciences geosoft-earthcube.org
- Geoscience Papers of the Future (GPF)
- Science is Changing, opening to community
- Research resources are becoming accepted as valuable contribution to science - Nature survey
- Software is valuable to scientist so they can do science instead of program, visualize, etc
- Sharing software methods needs to be improved
- Many reasons scientists don't want to share software.
- "Dark Software" is software lost rather than archived.
- Goal of project is to train geoscientists on how to share software.
- Software registry available: geosoft.earthcube.org
- New future vision: data published with paper
- Data: public repo: data, metadata, license, citable with unique and persistent identifier
- Software: public repo, doc, license, identifier
- Provenance described in workflow
- Call for papers: Earth and Science Science open July 1, 2015 to Jan 1, 2016 tinyurl.com/ess-gpf
- Repo on Github, PID from Zenodo
- Citation of software: Creator. (Year) Title. Repository. Version. PID. Date of retrieval.

- Questions -

- Papers are really just metadata about data and software. GPF isn't really changing what a paper looks like, but just formalizing code sharing.
    - Code can be in something like a python notebook.
    - GPF hopes to extend to other papers.
    - Want to improve reproducibility of scientific studies, want to help reader as much as possible.
    - Also provide workflow from downloading data to visualization.
    - Licensing software, how is it done?
- Many issues with licensing of software, especially if not open source.
    - GPF is required to be open source.

Jessica Hausman - Discussion on Citing Software and Subsetted Data

- Discussion of topics
- How do you handle versioning of online tools where frontend changes but backend doesn't, etc
- What parameters, flags, user inputs, etc should be included

- Why are we always (only?) considering DOIs? There are many other possibilities.
    - PURL, ARK, URN, UUID
- Citation needs to be understandable and usable by scientific community.
- Logs could be maintained that are linked to for versioning and dataset subsets.

- It seems there is not a generic approach that can apply to all cases of software or subsets.
    - Recommended way could come from provider.

- Mechanism discussed in Dynamic Citations Workshop.
    - Dataset citation separated from subset, so you have two identifiers.
    - DOI (or whatever) for data set, query identifier
    - Query ID must be maintained by repository. Is that a worthy effort for the repo?
    - Works well for long tail data, SQL databases, spreadsheets.
    - Worth and demand can scale with amount of queries.
    - Limited funds is a concern for this effort.

- Reproducibility has become an issue. You have to describe data used in study, and it shouldn't be cherry picked.

- Services to shorten unique identifiers and keep them persistent could take the burden off of repos.
    - Perhaps DataCite or publishers. Repos job is to archive and steward data.
    - Would have same business issues as repos now, money stream.

- Really long query strings becomes less of a problem when publications are not physical.
    - Some publications are becoming more multimedia, with embedded videos, etc.
    - Software could be included in such a publication.

- Group: COPDES about publishers and repositories working together to driven citations

- What about reproducing system environment to reproduce study?

- Brokers can get in the middle and reformat data that may not be logged completely.
    - Used to have Unique ID for each step. Now Provenance might not be recorded, especially for virtual steps.
    - Levels of products are cited, and each level references the level before and provenance in between.

- Books, regular citations, only go back one step (usually). Good analog for datasets.

Session Leads:

Name: Jessica Hausman [13]
Organization(s): NASA JPL PO.DAAC [14]
Email: Jessica.K.Hausman@jpl.nasa.gov [15]

Name: Daine Wright [16]
Organization(s): ORNL DAAC [17]
Email: wrightdm@ornl.gov [18]

Name: Downs, Robert R. [19]
Organization(s): Columbia University [20]
Email: rdowns@ciesin.columbia.edu [21]

Notes takers:

Name: Daine Wright [16]
Organization(s): ORNL DAAC [17]
Email: wrightdm@ornl.gov [18]

Participants:
Bob Downs, Ji-Hyun Oh, Shannon Rauch, Heather Brown, John Relph, Rebecca Fowler, Ward Flevri, Doug Schuster, Tiffany Mathews, Denise Hillls, Sarah Ramdeen, Nic Weber, Sandra Cosic, Steve Kempler, Wade Bishop, Tom Narock, Cyndy Chandler, Greg Janee, Audrey Mickle, Jamie Ryan, Ruth Duerr, Steve Olding, Madison Langseth, Daine Wright, Jessica Hausman

Accepted:
Keywords: Citation [22]

Source URL: https://commons.esipfed.org/node/7980

Links
[1] https://commons.esipfed.org/node/7980
[2] https://commons.esipfed.org/2015SummerMeeting
[3] https://commons.esipfed.org/session-type/breakout
[4] https://commons.esipfed.org/taxonomy/term/260
[5] https://commons.esipfed.org/collaboration-area/information-quality
[6] https://commons.esipfed.org/collaboration-area/preservation-and-stewardship
[7] https://commons.esipfed.org/collaboration-area/products-and-services
[8] https://commons.esipfed.org/taxonomy/term/1310
[9] https://goo.gl/0ZkxEH
[10] http://slides.com/shannonr/supporting-citation-of-subsetted-data-a-data-center-s-perspective
[11] http://geosoft-earthcube.org/gpf/
[12] http://onlinelibrary.wiley.com/journal/10.1002/%28ISSN%292333-5084/homepage/call_for_papers.htm
[13] https://commons.esipfed.org/node/7979
[14] https://commons.esipfed.org/taxonomy/term/408
[15] mailto:Jessica.K.Hausman@jpl.nasa.gov
[16] https://commons.esipfed.org/node/2603
[17] https://commons.esipfed.org/taxonomy/term/239
[18] mailto:wrightdm@ornl.gov
[19] https://commons.esipfed.org/node/1922
[20] https://commons.esipfed.org/taxonomy/term/234
[21] mailto:rdowns@ciesin.columbia.edu
[22] https://commons.esipfed.org/taxonomy/term/1882