TetherlessWorld

Rensselaer

Writing an article with excellent supporting data

Mark A. Parsons— 0000-0002-7723-0950 — @chutneyboy

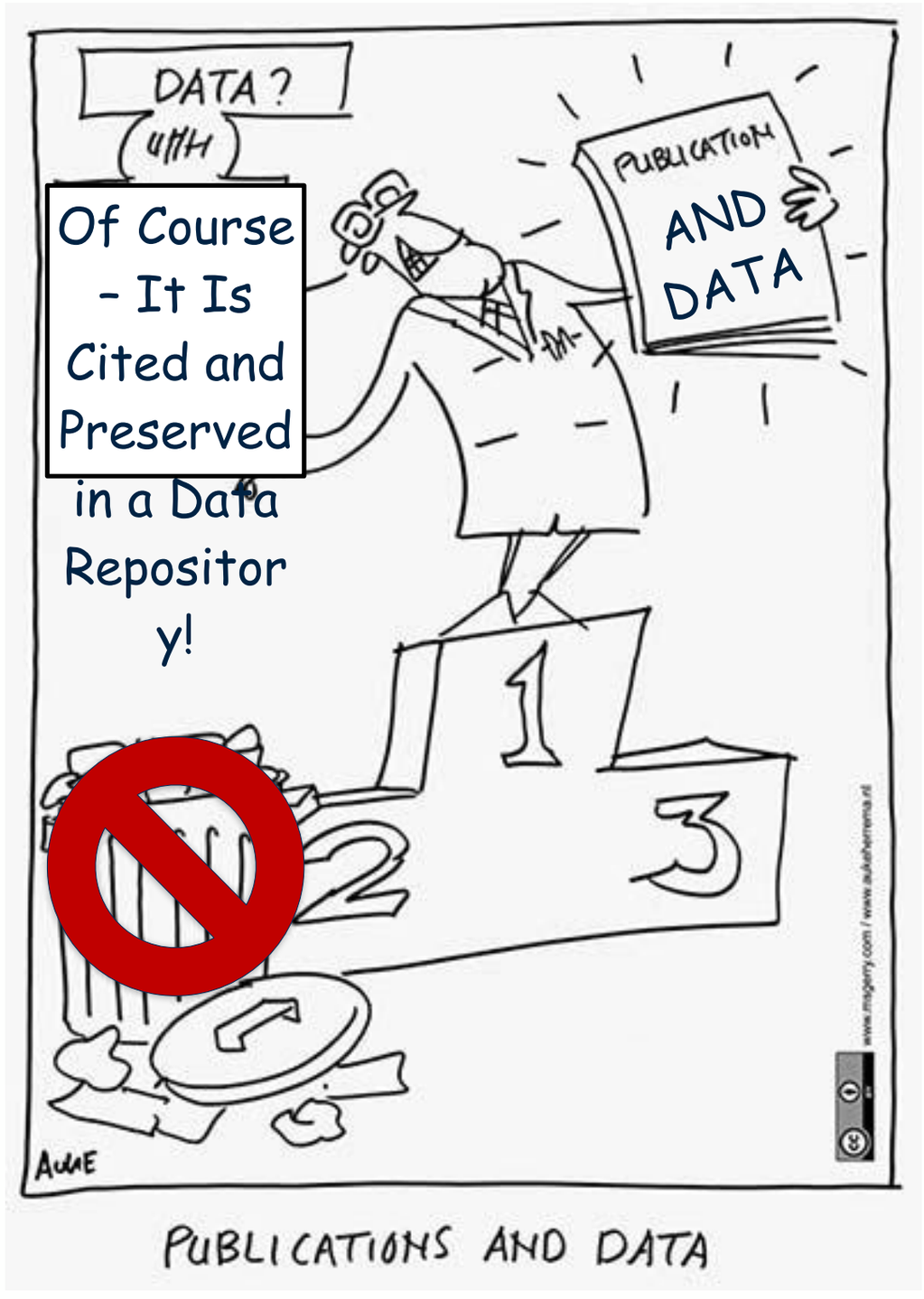
CODATA Connect Research Skills Development Webinar
22 June 2020

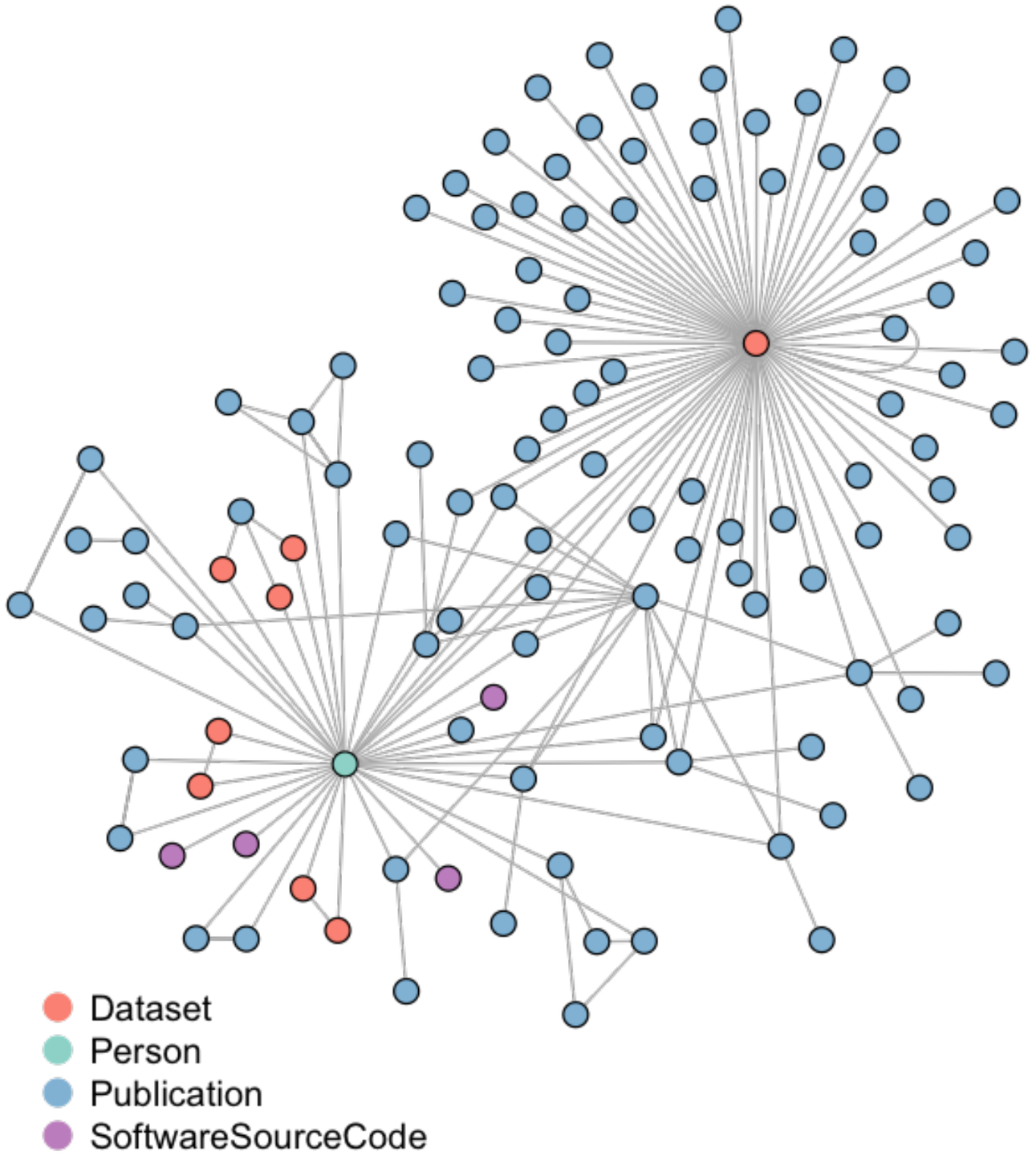
1st Class Research Product

RDA Plenary Cartoons by Auke Herrema.

1st Class Research Product

RDA Plenary Cartoons by Auke Herrema.





Research artifacts of a scholar

Fenner, M. (2019, October 12). FREYA PID Graph for a specific researcher (Version 1.0.0). DataCite. https://doi.org/10.14454/628M-3882

Rensselaer

Transparency and Openness Promotion (TOP) Guidelines for Journal Policies and Practice

TetherlessWorld

Citation Standards	Article is not published until providing appropriate citation for data and materials following journal's author guidelines.
Data Transparency	Data must be posted to a trusted repository, and reported analyses will be reproduced independently prior to publication.
Analytic Methods (Code) Transparency	Code must be posted to a trusted repository, and reported analyses will be reproduced independently prior to publication.
Research Materials Transparency	Materials must be posted to a trusted repository, and reported analyses will be reproduced independently prior to publication.
Design and Analysis Transparency	Journal requires and enforces adherence to design transparency standards for review and publication.
Study Preregistration	Journal requires preregistration of studies and provides link and badge in article to meeting requirements.
Analysis Plan Preregistration	Journal requires preregistration of studies with analysis plans and provides link and badge in article to meeting requirements.

https://www.cos.io/our-services/top-guidelines

Rensselaer

Transparency and Openness Promotion (TOP) Guidelines for Journal Policies and Practice

TetherlessWorld

Citation Standards	Article is not published until providing appropriate citation for data and materials following journal's author guidelines.
Data Transparency	Data must be posted to a trusted repository, and reported analyses will be reproduced independently prior to publication.
Analytic Methods (Code) Transparency	Code must be posted to a trusted repository, and reported analyses will be reproduced independently prior to publication.
Research Materials Transparency	Materials must be posted to a trusted repository, and reported analyses will be reproduced independently prior to publication.
Design and Analysis Transparency	Journal requires and enforces adherence to design transparency standards for review and publication.
Study Preregistration	Journal requires preregistration of studies and provides link and badge in article to meeting requirements.
Analysis Plan Preregistration	Journal requires preregistration of studies with analysis plans and provides link and badge in article to meeting requirements.

https://www.cos.io/our-services/top-guidelines

Rensselaer

Peter Brewer, EOS Article, September 14, 2017
"Do you expect me to just give away my data?"

- Dr. Brewer is an ocean chemist, and Senior Scientist, at the Monterey Bay Aquarium Research Institute (MBARI).
- At MBARI he served as President and Chief Executive Officer from 1991-1996
- Editor-in-Chief of AGU's *JGR: Oceans*
- Read more: https://www.mbari.org/brewer-peter/

The next six slides come from Stall, Shelley, Townsend, Randy, & Robinson, Erin. (2020, April). The Paper and The Data: Authors, Reviewers, and Editors Webinar on Updated Journal Practices for Data (and Software). Zenodo. http://doi.org/10.5281/zenodo.3744660
A four module webinar and slide deck that is an an excellent resource for more on this topic

Necessity and Benefits of data sharing…

"As Editor-in-Chief of a major AGU journal I frequently deal with inquiries from authors about the process of having their papers accepted for publication. Nothing in the recent past has generated more correspondence than the matter of data reporting requirements, particularly the fact that we no longer permit the statement of "Data available by contacting the author." I would like to describe, using some personal examples, why AGU's new data policy is both necessary and a benefit to all."

Brewer, P. (2017), "Do you expect me to just give away my data?", *Eos, 98,* https://doi.org/10.1029/2018EO081175. Published on 14 September 2017.

UK Royal Society Discussion Meeting on *Ocean Ventilation and Deoxygenation in a Warming World*.

"In order to do this, I wrote to the authors of the papers asking for copies of the data files accompanying their published work.

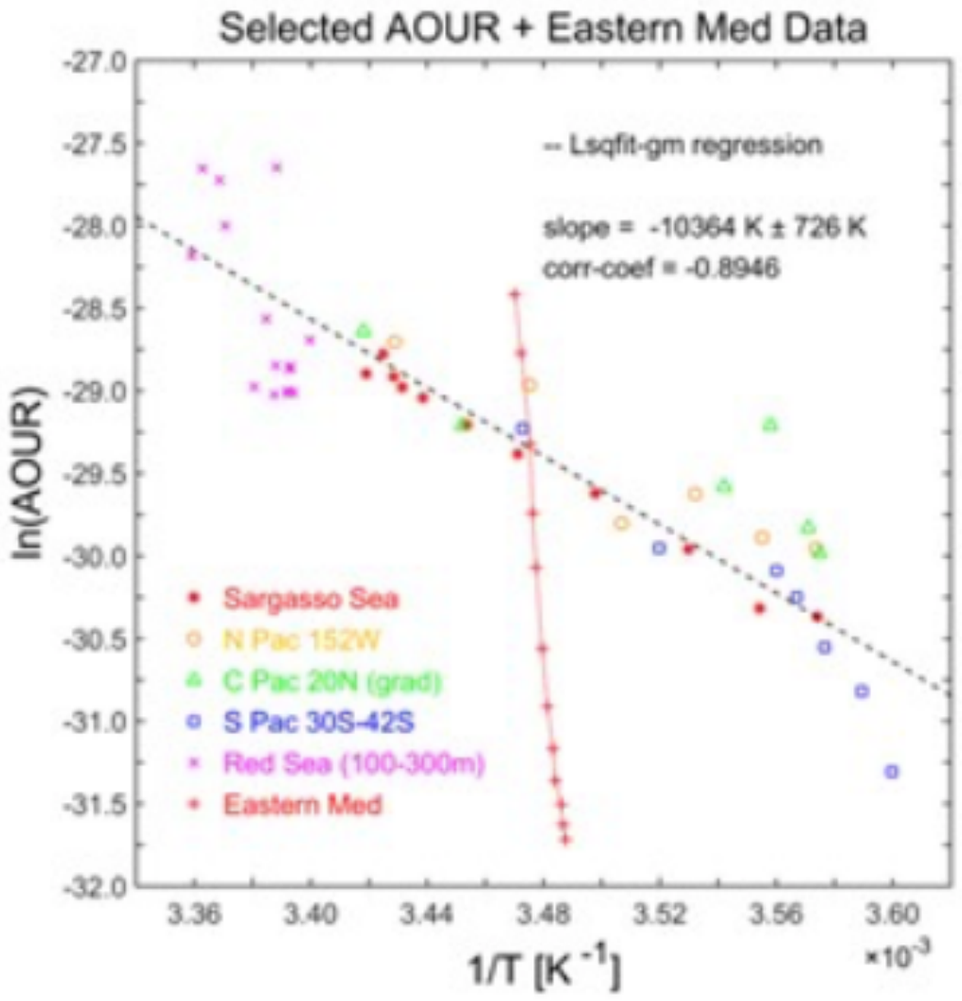
Not once did I receive a "Sure, I am happy to do that" reply: delay, obfuscation, pleas of being busy, and so on was the rule.

In many cases these were colleagues I knew well.

These are not big data files and typical profiles are only of a dozen or so depths. These are simple requests, but help was not at hand."

Brewer, P. (2017), "Do you expect me to just give away my data?", *Eos, 98,* https://doi.org/10.1029/2018EO081175. Published on 14 September 2017.

...one ocean profile ran almost exactly orthogonal to all the others...



Alas, too much time had gone by since the paper was published in 2001 and our colleague was now suffering from badly failing eye sight, the student who executed the work and provided the data equation had moved on and contact had been lost. The original data are now not discoverable and we have no way of resolving this puzzle.

Brewer, P. (2017), "Do you expect me to just give away my data?", *Eos, 98,* https://doi.org/10.1029/2018EO081175. Published on 14 September 2017.

...all the numbers...

"When you publish a research paper, you are also simultaneously publishing the data that supports your work. The readers of your article have equal rights to see both the words and the numbers – they are inseparable."

"We need the numbers – all the numbers – behind the published figures, graphs, contour plots etc. And these need to be specific; that is not averages within regions and so on."

Brewer, P. (2017), "Do you expect me to just give away my data?", *Eos, 98,* https://doi.org/10.1029/2018EO081175. Published on 14 September 2017.

My experience has been sobering.

"We need to do better than this.

Your data are valuable and need to be shared – this will bring you nothing but credit and honor."

Brewer, P. (2017), "Do you expect me to just give away my data?", *Eos, 98,* https://doi.org/10.1029/2018EO081175. Published on 14 September 2017.

What data?

- *Any* data you used in the research, including
  - Existing data
  - New data collected or created for this research

What to do with existing data

TetherlessWorld

- Formally cite any data you use and list them in the reference list.
- Reference any subset used.
- Include a data availability statement in the manuscript outlining any access requirements or restrictions
- Explain what you did with the data in the methods section, including what parts and versions were used
- Work with the repository to start thinking about how to reference subsets and other types of dynamic data.

Rensselaer

The Noble Eight-Fold Path to Citing Data

1. Importance
2. Credit and attribution
3. Evidence
4. Unique Identification
5. Access
6. Persistence
7. Specificity and verifiability
8. Interoperability and flexibility

Data Citation Synthesis Group. 2014. “Joint declaration of data citation principles.” https://dx.doi.org/10.25490/a97f-egyk.

Basic data citation form and content

TetherlessWorld

Per DataCite:

Creator. PublicationYear. Title. [Version]. Publisher. ResourceType. Identifier.

Per ESIP:

Author(s). Public Release Date. Title, Version ID. Repository. Resolvable Persistent Identifier. [date/time accessed]. [subset used].

Rensselaer

Author or Creator
The people or organizations responsible for the intellectual work to develop a data set. The data creator.

Title
The formal title of the data set. It may also include version or edition information but should be carefully controlled. A better alternative is to track version information independent of the title. Note this is the title of the data set, not the project or a related publication. It is important for the data set to have an identity and title of its own.

Public Release Date
When the particular version of the data set was first made available for use (and potential citation) by others.

Cline, D., R. Armstrong, R. Davis, K. Elder, and G. Liston. 2003. CLPX-Ground: ISA snow depth transects and related measurements ver. 2.0. Edited by M. A. Parsons and M. J. Brodzik. NASA National Snow and Ice Data Center Distributed Active Archive Center. https://doi.org/10.5060/D4MW2F23. Accessed 2008-05-14.

Version ID
Careful versioning and documentation of version changes are central to enabling accurate citation. Data stewards need to track and clearly indicate precise versions as part of the citation for any version greater than 1.

Resolvable Persistent Identifier
The unique identifier that provides the ability to access the data. Not all data have Persistent Identifiers (PIDs) or can be digitally accessed, so an alternative method to access metadata, such as a URL or a physical address, can be provided instead.

Repository
The name of the entity that holds, archives, publishes, prints, distributes, releases, issues, or produces the data. This may be an appropriate place to recognize a major sponsor of the data.

Access Date
Because data can be dynamic and changeable in ways that are not always reflected in release dates and versions, it is important to indicate when online data were accessed.

Author or Creator
The people or organizations responsible for the intellectual work to develop a data set. The data creator.

Title
The formal title of the data set. It may also include version or edition information but should be carefully controlled. A better alternative is to track version information independent of the title. Note this is the title of the data set, not the project or a related publication. It is important for the data set to have an identity and title of its own.

Public Release Date
When the particular version of the data set was first made available for use (and potential citation) by others.

Cline, D., R. Armstrong, R. Davis, K. Elder, and G. Liston. 2003. CLPX-Ground: ISA snow depth transects and related measurements ver. 2.0. Edited by M. A. Parsons and M. J. Brodzik. NASA National Snow and Ice Data Center Distributed Active Archive Center. https://doi.org/10.5060/D4MW2F23. Accessed 2008-05-14.

Version ID
Careful versioning and documentation of version changes are central to enabling accurate citation. Data stewards need to track and clearly indicate precise versions as part of the citation for any version greater than 1.

Resolvable Persistent Identifier
The unique identifier that provides the ability to access the data. Not all data have Persistent Identifiers (PIDs) or can be digitally accessed, so an alternative method to access metadata, such as a URL or a physical address, can be provided instead.

Repository
The name of the entity that holds, archives, publishes, prints, distributes, releases, issues, or produces the data. This may be an appropriate place to recognize a major sponsor of the data.

Access Date
Because data can be dynamic and
or more specific identifier
is important to indicate when online data were accessed.

Subset used
A crude way to identify the particular data used in a study

Hall, D. K. and G. A. Riggs. 2016. MODIS/Terra Snow Cover Daily L3 Global 500m Grid, Version 6. Oct. 2007- Sep. 2008, 84°N, 75°W; 44°N, 10°W. NASA National Snow and Ice Data Center Distributed Active Archive Center. https://doi.org/10.5067/MODIS/MOD10A1.006. Accessed 2019-02-02.

Subset used

or more specific identifier

study

Hall, D. K. and G. A. Riggs. 2016. MODIS/Terra Snow Cover Daily L3 Global 500m Grid, Version 6. Oct. 2007- Sep. 2008, 84°N, 75°W; 44°N, 10°W. NASA National Snow and Ice Data Center Distributed Active Archive Center. https://doi.org/10.5067/MODIS/MOD10A1.006. Accessed 2019-02-02.

Online example:

https://doi.org/10.1007/s11242-018-1079-1

Data Citation references

- Ball, A. & Duke, M. 2015. “How to cite datasets and link to publications.” DCC How-to Guides. Edinburgh: Digital Curation Centre. Available at http://www.dcc.ac.uk/resources/how-guides. Accessed 2018-02-10.
- CODATA/ICSTI Task Force on Data Citation. 2013. “Out of cite, out of mind: The current state of practice, policy and technology for data citation.” Data Science Journal 12: 1-75., http://dx.doi.org/10.2481/dsj.OSOM13-043.
- Data Citation Synthesis Group. 2014. Joint Declaration of Data Citation Principles. Martone M. (ed.). Force11. https://doi.org/10.25490/a97f-egyk.
- DataCite Metadata Working Group. 2017. DataCite Metadata Schema Documentation for the Publication and Citation of Research Data. Version 4.1. http://dx.doi.org/10.5438/0014.
- ESIP Data Preservation and Stewardship Committee. 2019. Data Citation Guidelines for Earth Science Data, Version 2. Earth Science Information Partners. https://doi.org/10.6084/m9.figshare.8441816.v1.
- Parsons, M. A., R. E. Duerr, and M. B. Jones. 2019. “The history and future of data citation in practice.” Data Science Journal 18 https://dx.doi.org/10.5334/dsj-2019-052.
- Rauber, A., A. Asmi, D. van Uytvanck, and S. Pröll. 2016. “Identification of reproducible subsets for data citation, sharing and re-use.” Bulletin of IEEE Technical Committee on Digital Libraries 12 (1): https://www.ieee-tcdl.org/Bulletin/v12n1/papers/IEEE-TCDL-DC-2016_paper_1.pdf. Accessed 2018-02-10.
- Stall, S., P. Cruse, H. Cousijn, J. Cutcher-Gershenfeld, A. de Waard, B. Hanson, J. Heber, K. Lehnert, M. Parsons, E. Robinson, M. Witt, L. Wyborn, L. Yarmey. (2018), Data sharing and citations: New author guidelines promoting open and FAIR data in the Earth, space, and environmental sciences, Sci. Editor 2018;41(3):83-87. https://www.csescienceeditor.org/article/data-sharing-and-citations-new-author-guidelines-promoting-open-and-fair-data-in-the-earth-space-and-environmental-sciences/. Accessed 2018-02-10.

What to do with newly created data

TetherlessWorld

All of the above, plus

- deposit the data in a trusted repository
- make data as open as possible and closed as ethically necessary
- good data management

Rensselaer

Choosing a repository — why?

TetherlessWorld

- **Improves discovery and citation of your data**
  - Data are assigned a unique, resolvable identifier (like a DOI), which allows unambiguous reference and discovery
  - Data are available in appropriate discovery portals etc.
- **Increases the reuse and hence value of your data**
  - community standard adherence
  - increasing usability
- **Makes your life easier**
  - Appropriate licensing and managed access.
  - User support
  - Guidance on all the publisher, funder, and metadata requirements

Choosing a repository — what type?

TetherlessWorld

- **Domain repository** — Repository specific to a a particular discipline or data type. Professional data stewards receive your data and help you ensure it align with community standards and practices.
    - Examples: CCDC, Protein Data Bank, PANGAEA
- **Thematic Repository** — similar to a domain repository but often provides additional services such as visualization or data integration around some scientific theme. May be a subset of a larger domain repository.
    - Examples: ELOKA, GenBank, CUAHSI
- **General repository** — all-purpose repositories which provide basic preservation, discovery, and reference services, but do not typically provide extensive curation or special services. Typically used for smaller data that do not have a domain repository.
    - Examples Zenodo, Figshare, institutional repositories

Choosing a repository — how?

TetherlessWorld

- Ask your funder. If they don't have an answer, encourage them to find an answer. It is their responsibility to support the stewardship of data they fund to collect.
- Ask your colleagues on what are the trusted repositories in your community.
- Ask your institution (library or research computing facility) if they have archiving services or policies. If they don't, encourage them to develop them.
- Use the DataCite data repository finder tool https://repositoryfinder.datacite.org. Favor certified trusted repositories, e.g. Core Trust Seal
- If all else fails, use Zenodo.

Rensselaer

Data Management Basics

TetherlessWorld

1st Principle: Data are for computers as much as humans. Good data are machine actionable.

1. Avoid manual processing. Use a scripted program for analysis like R, Python (Jupyter Notebooks), MATLAB, etc.
2. Store data in nonproprietary formats.
3. Store the raw, uncorrected data and record any changes to new versions (see 1).
4. Back up your data, i.e. use the cloud.
5. Use naming conventions. Be consistent and descriptive in data file names, table headings etc. Use full length standardized dates (e.g., 2020-06-22). Define your variables.
6. Keep your data 'tidy' and minimize special characters (Stick to ASCII when possible). Only record things once.
7. Make you data as self-describing as possible, at least include header rows. Don't mix data types (text, numeric) in a cell.
8. Document everything *especially uncertainty,* and use standard metadata templates.
9. Link your data to your publications.
10. **Use and support well-curated repositories! They help with all of the above.**

"Fold knowledge into data so program logic can be stupid and robust."

— Eric Raymond. 2004. *The Art of Unix Programming*

Rensselaer

A STORY TOLD IN FILE NAMES:

Location: C:\user\research\data

Filename	Date Modified	Size	Type
data_2010.05.28_test.dat	3:37 PM 5/28/2010	420 KB	DAT file
data_2010.05.28_re-test.dat	4:29 PM 5/28/2010	421 KB	DAT file
data_2010.05.28_re-re-test.dat	5:43 PM 5/28/2010	420 KB	DAT file
data_2010.05.28_calibrate.dat	7:17 PM 5/28/2010	1,256 KB	DAT file
data_2010.05.28_huh??.dat	7:20 PM 5/28/2010	30 KB	DAT file
data_2010.05.28_WTF.dat	9:58 PM 5/28/2010	30 KB	DAT file
data_2010.05.29_aaarrrgh.dat	12:37 AM 5/29/2010	30 KB	DAT file
data_2010.05.29_#$@*&!!.dat	2:40 AM 5/29/2010	0 KB	DAT file
data_2010.05.29_crap.dat	3:22 AM 5/29/2010	437 KB	DAT file
data_2010.05.29_notbad.dat	4:16 AM 5/29/2010	670 KB	DAT file
data_2010.05.29_woohoo!!.dat	4:47 AM 5/29/2010	1,349 KB	DAT file
data_2010.05.29_USETHISONE.dat	5:08 AM 5/29/2010	2,894 KB	DAT file
analysis_graphs.xls	7:13 AM 5/29/2010	455 KB	XLS file
ThesisOutline!.doc	7:26 AM 5/29/2010	38 KB	DOC file
Notes_Meeting_with_ProfSmith.txt	11:38 AM 5/29/2010	1,673 KB	TXT file
JUNK...	2:45 PM 5/29/2010		Folder
data_2010.05.30_startingover.dat	8:37 AM 5/30/2010	420 KB	DAT file

Type: Ph.D Thesis Modified: too many times

Data Management Basics

TetherlessWorld

1st Principle: Data are for computers as much as humans. Good data are machine actionable.

1. Avoid manual processing. Use a scripted program for analysis like R, Python (Jupyter Notebooks), MATLAB, etc.
2. Store data in nonproprietary formats.
3. Store the raw, uncorrected data and record any changes to new versions (see 1).
4. Back up your data, i.e. use the cloud.
5. Use naming conventions. Be consistent and descriptive in data file names, table headings etc. Use full length standardized dates (e.g., 2020-06-22). Define your variables.
6. Keep your data 'tidy' and minimize special characters (Stick to ASCII when possible). Only record things once.
7. Make you data as self-describing as possible, at least include header rows. Don't mix data types (text, numeric) in a cell.
8. Document everything *especially uncertainty,* and use standard metadata templates.
9. Link your data to your publications.
10. **Use and support well-curated repositories! They help with all of the above.**

"Fold knowledge into data so program logic can be stupid and robust."

— Eric Raymond. 2004. *The Art of Unix Programming*

Rensselaer

Data Management Guidance

TetherlessWorld

- Tidy Data by Hadley Wickham 10.18637/jss.v059.i10
  - “Tidy datasets are easy to manipulate, model and visualize, and have a specific structure: each variable is a column, each observation is a row, and each type of observational unit is a table.”
- Some Simple Guidelines for Effective Data Management by Elizabeth T. Borer et al. 10.1890/0012-9623-90.2.205
- Ten Simple Rules for the Care and Feeding of Scientific Data by Alyssa Goodman et al. 10.1371/journal.pcbi.1003542
- Support Your Data: A Research Data Management Guide for Researchers by John A Borghi et al. 10.3897/rio.4.e26439
- Data Management Training Clearinghouse http://dmtclearinghouse.esipfed.org
  - a favorite: “How to avoid a spreadsheet mess - Lessons learned from an ecologist”

A final note on PIDs

TetherlessWorld

- You cannot find, access, interoperate or reuse something unless you know what and where it is. This is especially true if you are a machine.
- We do this with persistent “actionable”* identifiers —PIDs: Unchanging names of entities (URNs) with a mechanism of resolving this to a location or access point — A Registry.
- DOIs the most known example, but there are many others

Rensselaer

Slide Credit: Martin Fenner, DataCite, PARSEC meeting at RDA P14, 21 Oct 2019

The FREYA project

Persistent Identifiers (PIDs) for the European Open Science Cloud (EOSC) with a focus on

- New PIDs (e.g. organizations, instruments)
- **PID Graph**: Connections between PIDs
- **PID Forum**: Training and Outreach
- **PID Commons**: Sustainability

Website: www.project-freya.eu

FREYA project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777523.

PID Graph is powered by GraphQL

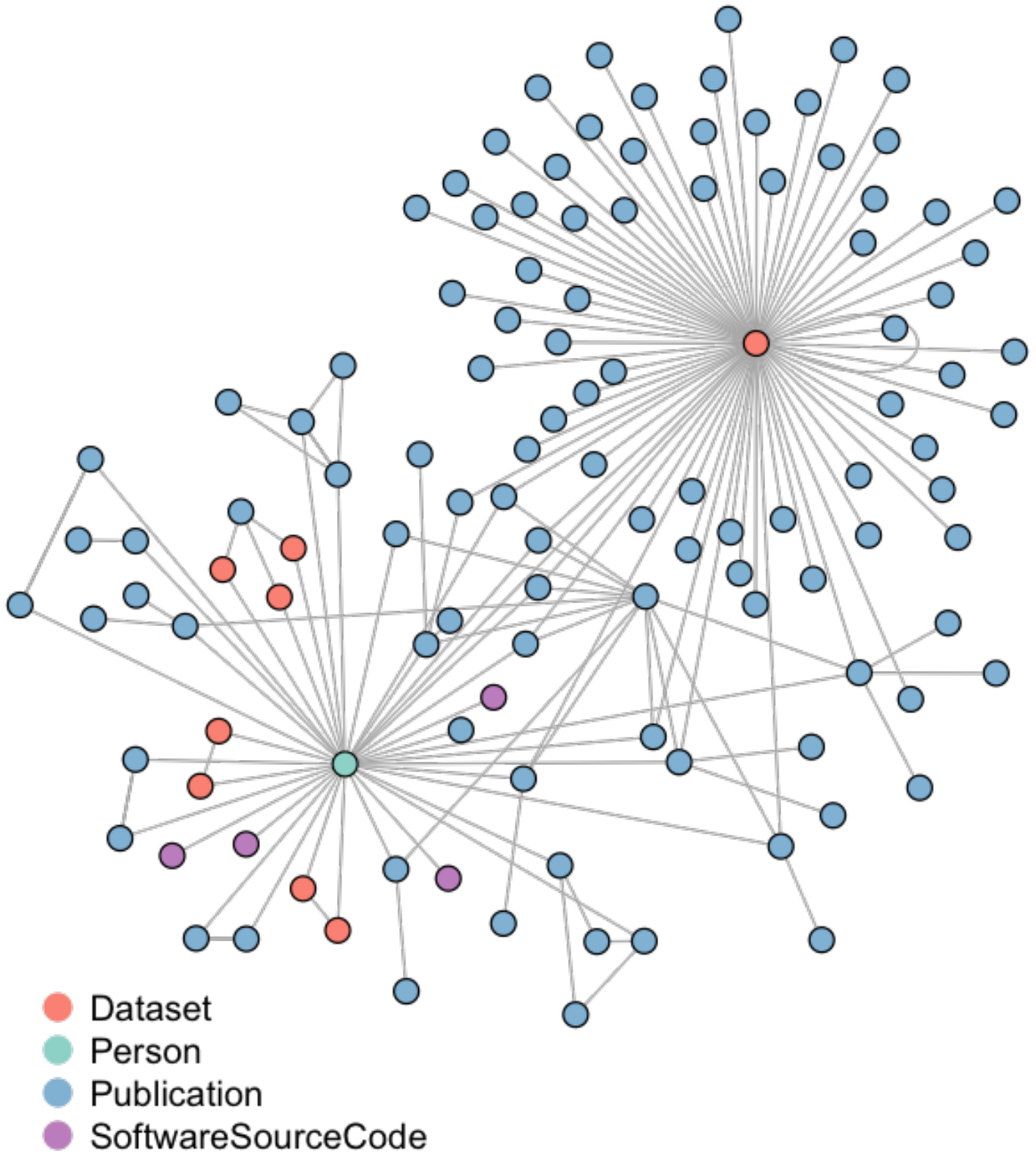
Slide Credit: Martin Fenner, DataCite, PARSEC meeting at RDA P14, 21 Oct 2019

https://api.datacite.org/graphql

PID Graph uses GraphQL to enable complex queries of the federated graph provided by FREYA partners

Jupyter Notebook

Fenner, M. (2019, June 30). FREYA PID Graph Key Performance Indicators (KPIs) (Version 1.1.0). DataCite. https://doi.org/10.14454/3BPW-W381



Research artifacts of a scholar

Fenner, M. (2019, October 12). FREYA PID Graph for a specific researcher (Version 1.0.0). DataCite. https://doi.org/10.14454/628M-3882

Rensselaer

ORCID example

TetherlessWorld

Rensselaer

Takeaways

- Cite data (and software)
- Work with a trusted repository and learn the basics of data stewardship (think about the machine)
- Get an ORCID and use it

Why should you care?

TetherlessWorld

- Credit! Papers with data citations and readily accessible data are cited more often. PLUS non-traditional citations capture what may be the more important artifacts.
- Learn how your data are being used and connect with possible collaborators
- Improve reproducibility and increase trust in science
- Comply with funder mandates
- Comply with publisher requirements

Rensselaer

TetherlessWorld

Rensselaer

Thank You!

Mark A. Parsons— 0000-0002-7723-0950 — @chutneyboy