COLUMBIA CLIMATE SCHOOL

CENTER FOR INTERNATIONAL EARTH SCIENCE INFORMATION NETWORK

Documenting Data to Improve Trust and Support Use Across Disciplines and Vocations

Robert R. Downs, PhD

Center for International Earth Science Information Network

Columbia Climate School, Columbia University

Prepared for presentation to the

47th IASSIST Annual Conference

7-10 June 2022 in Gothenburg, Sweden and Virtual

Session: C3: Data Documentation and Reproducibility, 8 June 2022

Abstract

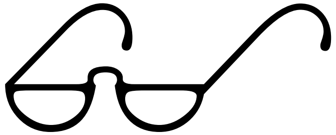
Data documentation allows audiences to determine whether and how to use data. Providing detailed descriptions of data products and services enables potential users to decide on the applicability of the data for an intended use. Furthermore, such data documentation also enables users to determine how the data could be used to meet a particular objective. Practically, creating effective data documentation requires detailed knowledge of the data that is ordinarily held by those who have been involved in the collection or production of the data. By utilizing such in-depth knowledge of a data product, rich documentation can be produced that is beneficial to data repository stakeholders. Offering rich data documentation provides opportunities to serve a diversified audience of users and potential users. The audience for data that includes rich documentation could span disciplinary and vocational boundaries, enabling use of data that is not necessarily dependent on the specialized knowledge that is associated with a specific discipline or sub-discipline. Providing rich documentation along with data also can improve the trust that the designated community has in a data repository by demonstrating the repository's commitment for enabling use. The challenge for creating rich data documentation is in obtaining the documentation from data producers who have not been developing or routinely preparing such documentation as part of their data collection practices. Likewise, data producers who do not expect to be recognized or rewarded for such contributions may not be sufficiently motivated to produce rich data documentation. Aspects of rich data documentation are described along with the challenges and benefits of providing such documentation with data.

Data Documentation: Important for Data Discovery and Exploration

- Data discovery: finding a data product
  - Are datasets available on the planned study topic?
  - Is data documentation available about the planned study topic?
- Data Exploration: deciding whether to consider using a data product
  - Is a particular dataset appropriate for the planned study?
  - Does the description of the data and variables match the research objective?

Data Documentation: Important for Research Decision-Making

- Data use: how a particular data product can be used
  - What methodology would be appropriate to apply to the data?
  - Are the results reproducible using the same methodology and data?
  - Are the results replicable using the same methodology to analyze other data?



- Data Interpretation: how to interpret the data
  - Can the data be interpreted in a manner that is consistent with the method?
- Data Integration: compatibility with other data
  - Can the data be combined with other data for the planned analysis?
  - Can the data be integrated with other data to support other analyses?
  - Can the data be included along with other data within a service?

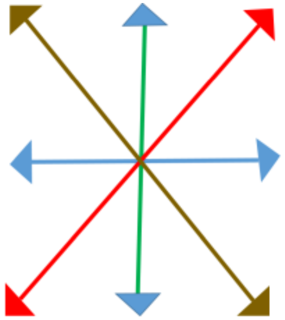
Data Documentation: Users Need to Determine Fitness for Use

- Are the methodological assumptions for data collection described?
  - Are the methods consistent with the proposed data use?
- Is the quality of the data described?
  - Is the data quality appropriate for the planned data use?
- Are the limitations of the data described?
  - Do the limitations prohibit the planned data use?

Value of Data Documentation: Data Use

- Enabling diversification of the user community:
  - Levels of expertise (senior scientists, early career and students, general public)
  - Interdisciplinary (studies within other fields, studies across fields)

- Demonstrates to stakeholders a commitment for enabling use
  - Data producers can see that their data deposit will facilitate data use
  - Data peer-reviewers have the ability to assess data
  - Users recognize the sources of data that are documented to support usage
  - Funders see that value produced by their investments
  - Publishers can justify recommending the repository for data producers to deposit their data
  - Establishes the data center as a model among data repositories



- All data are not the same in terms of value to users
  - Wide range of review practices among data repositories
  - Data documentation is a differentiating factor for data products

Data Documentation Addresses Current Data Principles

- FAIR Principles
  - Supporting findability, accessibility, interoperability, and reusability[1]
- CARE Principles
  - Fostering collective benefit, authority to control, responsibility, and ethics[2]
- TRUST Principles
  - Facilitating transparency, responsibility, user focus, sustainability, and technology[3]
- GEO Data Management Principles
  - Enabling discovery, access, use, preservation, and curation[4]

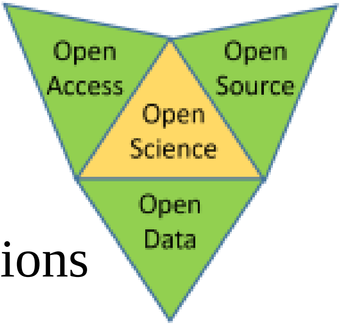
[1] Wilkinson, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

[2] Carroll, et al. 2020. The CARE Principles for Indigenous Data Governance. Data Science Journal, 19(1), p.43. DOI: http://doi.org/10.5334/dsj-2020-043

[3] Lin, D., Crabtree, J., Dillo, I. et al. The TRUST Principles for digital repositories. Sci Data 7, 144 (2020). https://doi.org/10.1038/s41597-020-0486-7

[4] Group on Earth Observations. Data Management Principles. https://earthobservations.org/open_eo_data.php#

Challenge: Recognizing Data Documentation as a Research Contribution



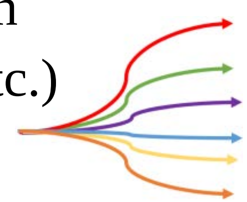
Open data documentation is a contribution to open science

  - Creating data documentation improves the usability and value of data
  - Documentation should be released with the data and freely available without restrictions

- Similarities between peer-reviewed data documentation and published data paper
  - Both should be publicly accessible and associated with the data
  - Both require similar effort and provide value to research data users
- Differences between peer-reviewed data documentation and published data paper
  - Focus on peer-review of data in conjunction with peer-review of documentation
  - Credit for peer-reviewed documentation versus credit for peer-reviewed data paper
- Hiring and promotion criteria must include open science research contributions
  - Publication of data, documentation, and software
  - Citation of data, documentation, and software

Data Repository Support for Data Documentation

- Encourage data producers to share data documentation
  - Describe the benefits of documenting data for enabling discovery and use
  - Link from data to data paper if available as an open access publication
- Assist data producers in preparing data documentation
  - Offer a data documentation template for sharing documentation in a consistent format
  - Prepare an initial draft based on data, metadata and communications with data producer
- Provide metrics for citations of data and data documentation
  - Display references to published articles that cited the data and data documentation
  - Display counts of publications (articles, books and book chapters, proceedings, etc.)



Helping Data Producers Prepare Data Documentation

Documentation for <Dataset Title>
<Documentation Publication Date>
<Authors>
Abstract
Data set citation
Suggested citation for documentation
Contact to provide feedback on documentation
Table of Contents
I. Introduction
II. Data and Methodology
III. Data Set Description(s)
IV. How to Use the Data
V. Potential Use Cases
VI. Limitations
VII. Acknowledgments
VIII. Disclaimer
IX. Use Constraints
X. Recommended Citation(s)
XI. Source Code
XII. References
XIII. Documentation Copyright & License
Appendix 1.
Data Revision History
Appendix 2.
Contributing Authors & Documentation Revision History

Source: Table of contents for the SEDAC Data Documentation Template

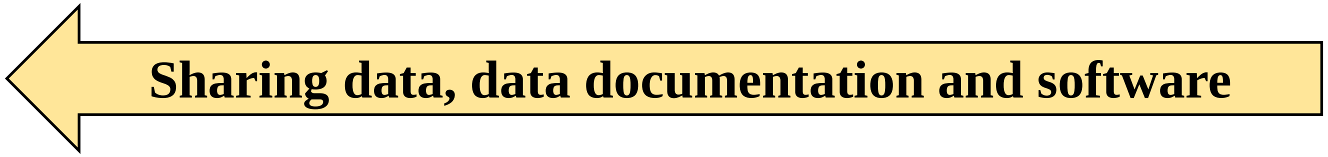
Motivating Data Documentation: Need to Change Research Culture

- Review data documentation as part of peer-review of data
  - Selective data repositories conduct rigorous review of data and documentation
- Recognize data documentation as a scholarly contribution
  - Reward the creation and sharing of data documentation as well as data and software
- Collect and reward metrics for data documentation
  - Consider shared data documentation as equivalent to a published data paper
  - Treat citations of data documentation like citations of a published paper

Overcoming Negative Perceptions of Data Documentation

Data documentation may be considered another task added to workload if not recognized!

- Traditional academic evaluation criteria
  - Teaching
  - Publications
  - Community Service



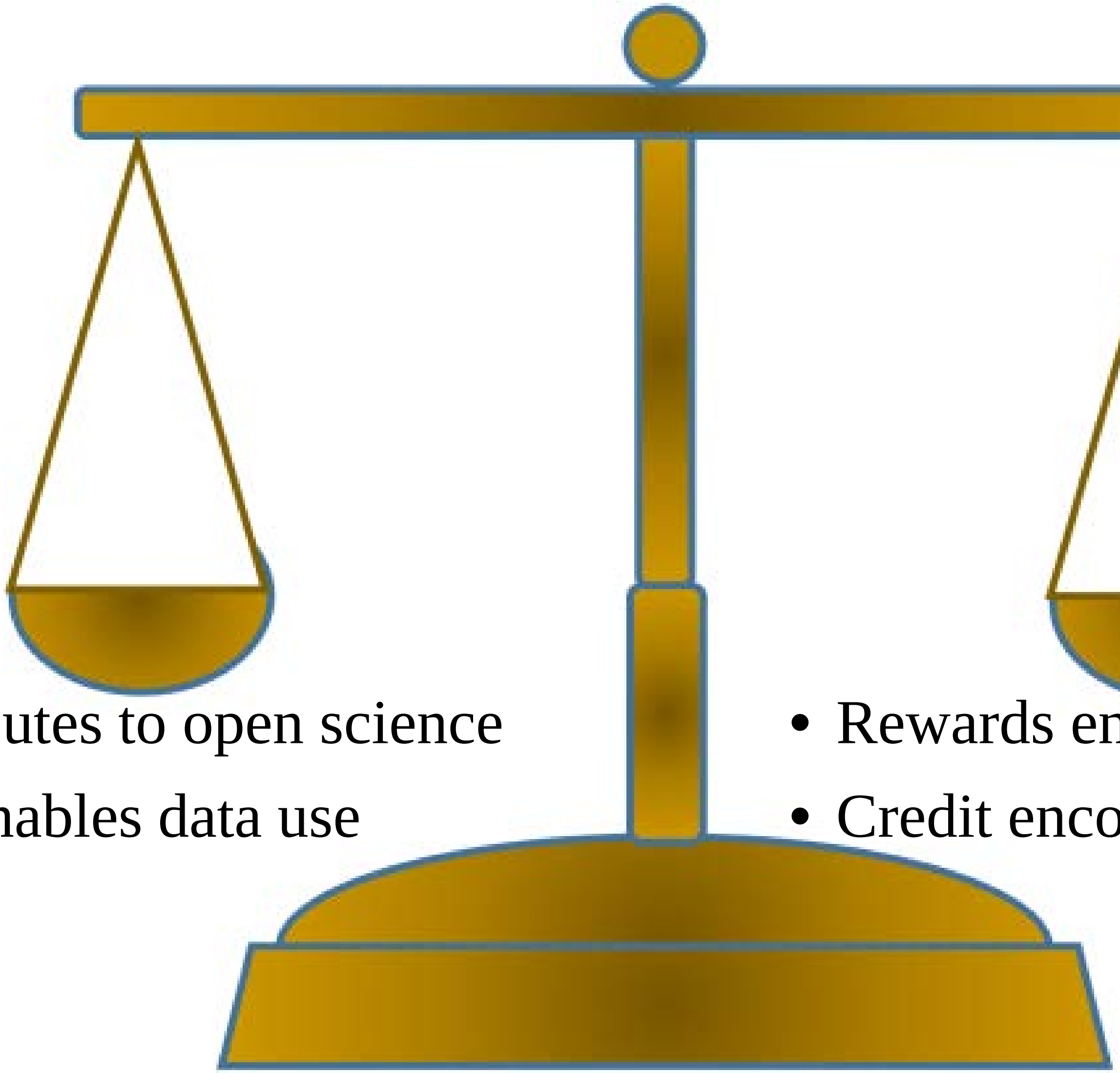
Sharing data, data documentation and software

Relevant DORA Recommendation for Funding Agencies and Institutions

"For the purposes of research assessment, consider the value and impact of all research outputs (including datasets and software) in addition to research publications, and consider a broad range of impact measures including qualitative indicators of research impact, such as influence on policy and practice" (DORA, 2013).

Source: San Francisco Declaration on Research Assessment (DORA) https://sfdora.org/read/

Balancing Research Contributions and Rewards for Open Science



- Open data contributes to open science
- Documentation enables data use

- Rewards encourage open science
- Credit encourages data documentation

Thank you for your interest!

rdowns@ciesin.columbia.edu