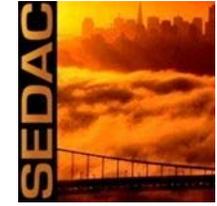


DOCUMENTATION TO FOSTER SHARING AND USE OF OPEN EARTH SCIENCE DATA: QUALITY INFORMATION



Robert R. Downs¹, Ge Peng², David F. Moroni³,
Hampapuram K. Ramapriyan⁴, Yaxing Wei⁵

¹Socioeconomic Data and Applications Center, Center for International Earth Science Information Network, Columbia University

²University of Alabama at Huntsville & NASA Marshall Space Flight Center, IMPACT Project

³Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA

⁴Science Systems and Applications, Inc. & NASA Goddard Space Flight Center

⁵Oak Ridge National Laboratory, Oak Ridge, TN

*Prepared for Presentation to the 16th International Digital Curation Conference
19 April 2021, Edinburgh, Scotland and Virtual
Session: FAIR and Metadata
15:00 - 15:40 BST.*



Outline of Topics

- Open Earth science data offers societal benefits
- Challenges for supporting data reuse by diverse users
- Common challenges for reusing data, including Earth science data
- Importance of dataset quality information - why should we care?
- Benefits of data quality information for users
- Notable challenges related to data quality information
- Development of guidelines for representing quality information
- How will guidance on quality information address the issue
- Plans and next steps

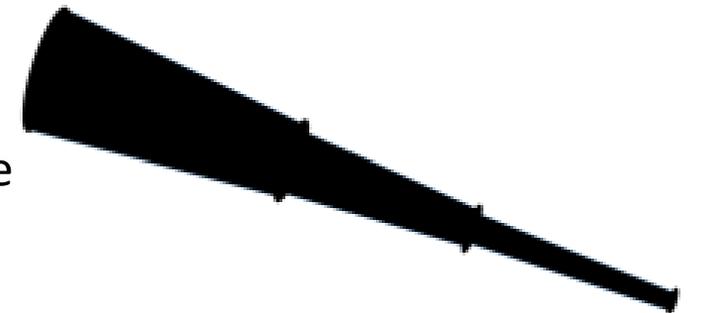
Open Earth Science Data Offers Societal Benefits

- Leverages data from previously conducted research
 - New value from previous investments
- Opportunities for new users to reuse data
 - Use beyond the data collection team
 - Use across disciplines and internationally
 - Use by broader communities, including practitioners and the general public
- New Opportunities for Data Reuse
 - New research hypotheses and questions
 - Development of new data products and services
 - Support for decision making
 - Knowledge hubs (digital infrastructure)
 - Government and commercial applications



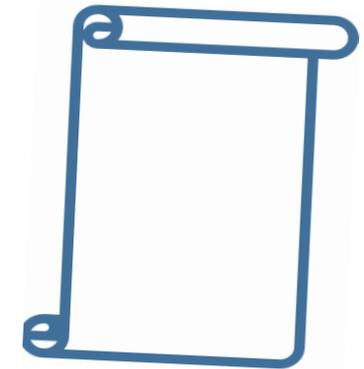
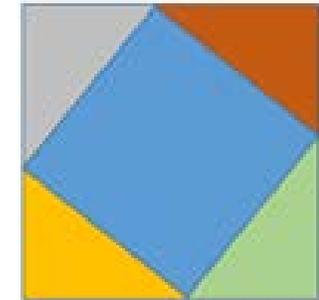
Selected Challenges for Supporting Data Reuse by Diverse Users

- Facilitate understanding by users
 - Packaged to foster access and ease of use
 - Described to facilitate quick discovery and selection
 - Documented to enable users to determine applicability for their objectives
 - Interoperability with other data products and services
 - Leverage and foster controlled vocabularies to facilitate consistent, semantic understanding
 - International use may require translations of documentation
 - Will the documentation provide enough context for future users to determine their applicability?
- Advance Open science
 - Clear and understandable data use policies that enable any use by anyone without restrictions
- Enable discovery, access, and use
 - Discoverable within a platform that can provide access and enable use
 - Harvested and indexed within catalogues that facilitate discovery
 - Capabilities for Data Reuse
 - Sustainable to continue supporting data archival, dissemination, and on-demand analysis
 - Data resilience - available for future use and as lasting evidence for decisions



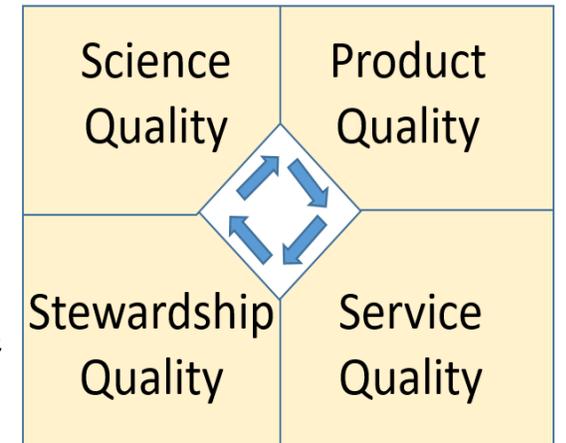
Common Challenges for Reusing Data, Including Earth Science Data

- Inconsistencies
 - Data descriptions (metadata and documentation) both within and across disciplines
 - Data structure, formats, variables
 - Tools and services for accessing and using data (and documentation on their use)
- Interoperability
 - Combining data produced from different methods
 - Enabling use in services and tools
 - Use for different applications, development of data products
 - Works toward ensuring data is analysis-ready
- Documentation of data quality is often missing
 - Support effective use
 - Determine applicability for a particular purpose
 - Enable informed and proper use
 - Efficient decision-making the appropriateness of data for planned use
 - Identify appropriate methods for analysis
 - Identify compatibility with other data products



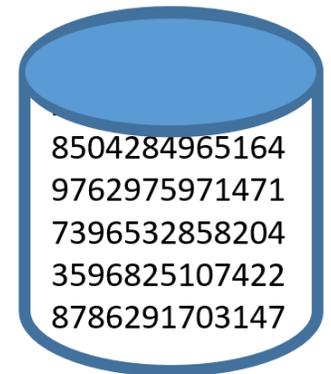
Importance of Dataset Quality Information - Why should we care?

- A lifecycle approach offers dimensions of dataset quality information:
 - science quality, product quality, stewardship quality, service quality
- Quality Information
 - Enables data reuse across communities
 - Reduces the potential for data misuse
 - Reuse scenario on data integration
 - Facilitates evaluation of data for a particular purpose
- Users need to be able to find answers to simple questions
 - Is the quality of the data acceptable?
 - Can the data be used with the method for the planned study?
 - Which data product is better in terms of quality?



Benefits of Data Quality Information for Users

- Facilitates the determination of the applicability of data
- Reduces the need to make assumptions about or evaluate data quality
- Enables informed use that facilitates decisions about proper use
- Allows more informed distinction among datasets containing similar measurements
- Fosters review of the data and of studies that used the data



Notable Challenges Related to Data Quality Information



- Data quality information is not discoverable
 - Users need to find data quality information when considering data
- Data quality information is not clearly described
 - Only understandable by data producers and scientists from that discipline
- Inconsistent practices
 - Collecting, preparing, representing, and sharing quality information

Proposed Solution: Develop Recommended Practices for Representing Quality Information as Advocated by Relevant Principles

- Findable, Accessible, Interoperable, and Reusable by applying the FAIR Principles¹ to data quality information that is provided in metadata and data documentation
- Assessed to facilitate Collective Benefit, Authority to Control, Responsibility, and Ethics in accordance with the CARE Principles for Indigenous Data Governance²
- The data repository should ensure that Transparency, Responsibility, User focus, Sustainability, and Technology are applied to quality information as per the TRUST Principles for Digital Repositories³
- Support open data by default with minimal restrictions or time delay as in the GEOSS Data Sharing Principles⁴
- Provide discoverability, accessibility, usability, preservation, and curation as described by the GEOSS Data Management Principles⁵

1. Wilkinson, et al. 2016. The FAIR Guiding Principles for Scientific Data Management and Stewardship. Sci Data 3. <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), DOI: <http://doi.org/10.5334/dsj-2020-043>

3. Lin, et al. 2020. The TRUST Principles for Digital Repositories. Scientific Data 7, 144. <https://doi.org/10.1038/s41597-020-0486-7>

4. GEOSS Data Sharing Principles. 2016. Group on Earth Observations. https://earthobservations.org/open_eo_data.php

5. GEOSS Data Management Principles. 2015. Group on Earth Observations. https://earthobservations.org/open_eo_data.php

How will guidance on quality information address the problem?

- Audience
 - Data producers, stewards, intermediaries, editors, funders
- Direction for improving practice
 - Reduce the learning curve for describing the quality of data
- Considerations for use across disciplines and skill levels
 - Advocate documentation and data tools features to support various skill levels
 - Exemplify quality information that employs limited jargon and defines discipline-specific terms
 - Recommend approaches for describing appropriate uses and limitations to avoid data misuse
 - Promote the use of clear language for non-scientists, students, practitioners, and the general public



How do we plan to do it? Next steps?



- Diversified volunteer contributors
 - Representing various domains
 - ESIP Information Quality Cluster (IQC) https://wiki.esipfed.org/Information_Quality
- Open access to guidelines
 - Publication under an open license, such as the Creative Commons Attribution 4.0 International (CC BY) license: <https://creativecommons.org/licenses/by/4.0/>
- Peng G, et al. 2021. International Community Guidelines for Sharing and Reusing Quality Information of Individual Earth Science Datasets. <https://doi.org/10.31219/osf.io/xsu4p>

