Documentation to Foster Sharing and Use of Open Earth Science Data: Quality Information

Robert R. Downs
Center for International Earth
Information Network (CIESIN),
Columbia University

Ge Peng
Earth System Science Center/NASA
MSFC IMPACT, The University of
Alabama in Huntsville

David F. Moroni
Jet Propulsion Laboratory,
California Institute of Technology,
Pasadena, CA

Hampapuram K. Ramapriyan
Science Systems and Applications,
Inc. and NASA Goddard Space Flight
Center

Yaxing Wei
Environmental Sciences Division,
Oak Ridge National Laboratory,
Oak Ridge, TN, 37830

Submitted 13 January 2021

Correspondence should be addressed to Dr. Robert R. Downs, CIESIN, The Earth Institute, Columbia University, 61 Route 9W, Palisades, NY USA 10964. Email: rdowns@ciesin.columbia.edu

The 16th International Digital Curation Conference takes place on 19 April 2021, Edinburgh, Scotland

URL: http://www.dcc.ac.uk/events/idcc2021

Providing capabilities to reuse open Earth science data offers opportunities to leverage observations from previously conducted research so that new research can be conducted. By supporting data producers in their efforts to share the data that they have collected, scientific data repositories enable broader audiences to reuse these data. However, simply providing access to data is not enough to facilitate understanding of the data by diverse users who did not collect the data and are not trained in the discipline represented by the data. Moreover, simply providing access to data is not enough to foster reuse by transdisciplinary users, practitioners, learners, and members of the general public who could benefit from reusing research data. If plans for the dissemination of data include reuse across disciplinary boundaries, including providing opportunities for non-scientific reuse, then the data must be packaged, described, and discoverable in a way that allows potential users to understand the data in terms of their own objectives. Effectively packaging, documenting, and indexing data offers opportunities to facilitate understanding of data and to foster reuse of data products and services by broader audiences as well as by those who are more familiar with the data. Furthermore, such data curation activities also can reduce the potential for misuse of the data.

What is often missing is the documentation of data quality that is needed to support the effective use of open data. In addition to ensuring that data are designated as open, the quality of datasets also must be described to foster more informed and proper usage. Information about the quality of open data should be clearly documented and discoverable such that it is available to all potential users of the data. When searching for open data to complete a project, potential users require information about the quality of the data to determine whether the quality of the data is sufficient to meet their data usage needs. Providing data quality information reduces the need to make assumptions when selecting among candidate data products and services for potential reuse. Similarly, data users require information on the data quality when deciding on the applicability of data and on the methods to be used for analyzing and interpreting the data. Information about the data quality should be clearly described so that it can be easily understood by potential users from cross-disciplinary fields, as well as by practitioners, such as operational forecasters, planners and decision makers.

Reuse scenarios demonstrate the need for providing understandable data quality information along with the data. For example, when identifying open data products for possible integration to create new data products or services, interoperable data integration workflows are developed to ensure that the quality of each candidate dataset is described. Data quality descriptions enable candidate data products to be assessed in terms of their compatibility with each other and their applicability for the purposes of the proposed data integration project. In these workflows, the quality of the resulting integrated dataset must be described in appropriately curated discovery metadata and in the data documentation to foster traceable, targeted, and efficient decision making about the potential of each data product and to support the data integration effort. The importance of documenting data quality is described and the value of effectively documenting and packaging data quality information is demonstrated with real-world research and operational use cases.