D

Data Sharing

Tao Wen
Earth and Environmental Systems Institute, Pennsylvania State University, University Park, PA, USA

Definition

In general, data sharing refers to the process of making data accessible to data users. It often happens through community-specific or general data repositories, personal and institutional websites, and/or data publications. A data repository is a place storing data and providing access to users. Data sharing is particularly encouraged in research communities although the extent to which data are being shared varies across scientific disciplines. Data sharing links data providers and users, and it benefits both parties through improving the reproducibility and visibility of research as well as promoting collaboration and fostering new science ideas. In particular, in the big data era, data sharing is particularly important as it makes big data research feasible by providing the essential constituent – data. To ensure effective data sharing, data providers should follow findability, accessibility, interoperability, and reusability (FAIR) principles (Wilkinson et al. 2016) throughout all stages of data management, a broader topic underpinned by data sharing.

FAIR Principles

Wilkinson et al. (2016) provide guidelines to help the research community to improve the findability, accessibility, interoperability, and reusability of scientific data. Based on FAIR principles, scientific data should be transformed into a machine-readable format, which becomes particularly important given that an enormous volume of data is being produced at an extremely high velocity. Among those four characteristics of FAIR data, reusability is the ultimate goal and the most rewarding step.

Findability

Data sharing starts with making the data findable to users. Both data and metadata should be made available. Metadata are used to provide information about one or more aspects of the data, e.g., who collect the data, the date/time of data collection, and topics of collected data. Each dataset should be registered and assigned a unique identifier such as a digital object identifier (DOI). Each DOI is a link redirecting data users to a webpage including the description and access of the associated dataset. Both data and metadata should be formatted following formal, open access, and widely endorsed data reporting standard (e.g., schema.org: https://schema.org/Dataset). Those datasets fulfilling these standards can be cataloged by emerging tools for searching datasets (e.g., Google Dataset Search: https://toolbox.google.com/datasetsearch). Currently, it is more common that data users will search for desired datasets

L. A. Schintler, C.L. McNeely (eds.), *Encyclopedia of Big Data*,
https://doi.org/10.1007/978-3-319-32001-4_322-1

through discipline-specific data repositories (e.g., EarthChem: https://www.earthchem.org/ in earth sciences).

Accessibility

Both data and metadata should be provided and can be transferred to data users through data repository. Broadly speaking, data repository can be personal- or institutional-level websites (e.g., Data Commons at Pennsylvania State University: http://www.datacommons.psu.edu) and discipline-specific or general databases (e.g., EarthChem). Data users should be able to use the unique identifier (e.g., DOI) to locate and access a dataset.

Interoperability

As more interdisciplinary projects are proposed and funded, shared data from two or more disciplines often need to be integrated for data visualization and analysis. To achieve interoperability, data and metadata should not only follow broadly adopted reporting standards but also use vocabularies to further formalize reported data. These vocabularies should also follow FAIR principles. The other way to improve interoperability is that data repositories should be designed to provide shared data in multiple formats, e.g., CSV and JavaScript Object Notation (JSON).

Reusability

Enabling data users to reuse shared data is the ultimate goal. Reusability is the natural outcome if data (and metadata) to be shared meet the rules mentioned above. Shared data can be reused for testing new science ideas or for reproducing published results along with the shared data.

The Rise of Data Sharing

Before the computer age, it was not uncommon that research data were published and deposited as paper copies. Transferring data to users often required individual request sent to the data provider. The development of the Internet connects everyone and allows data sharing almost in real time (Popkin 2019).

Nowadays more data are shared through a variety of data repositories providing access to data users. The scientific community including funders, publishers, and research institutions has started to promote the culture of data sharing and making data open access. For example, the National Science Foundation requires data management plans in which awardees need to describe how research data will be stored, published, and disseminated. Many publishers, like *Springer Nature*, also require authors to deposit their data in general or discipline-specific data repository. In addition to sharing data in larger data repositories funded by national or international agencies, many research institutions start to format and share their data in university-sponsored data repositories for the purpose of long-term data access.

In some disciplines, for example, Astronomy and Meteorology, where data collection often relies on large and expensive facilities (e.g., satellite, telescope, a network of monitoring stations) and the size of dataset is often larger than what one research group can analyze, data sharing is a common practice (Popkin 2019). In some other disciplines, some researchers might be reluctant to share data for varying reasons. These reasons can be in the processes of data publication and data citation. Some of these reasons include:

(1) Researchers are concerned that they might get scooped if they share data too early.
(2) Researchers might lack of essential expertise to format their data to certain standard.
(3) Funding that supports data sharing might not be available to these researchers to pay for their time to make data FAIR.
(4) The support for building data repositories is insufficient in some disciplines.
(5) The research community fails to treat data sharing as important as publishing journal article.
(6) Insufficient credit has been given to data providers as data citation might not be done appropriately by data users.

To address some of these problems, all stakeholders of data sharing are working

collaboratively. For example, European Union projects FOSTER Plus and OpenAIRE provide training opportunities to researchers on open data and data sharing. The emerging data journals, e.g., *Nature Scientific Data*, provide a platform for researchers to publish and share their data along with descriptions. Many funders, including the National Science Foundation, have allowed repository fees on grants (Popkin 2019).

Best Practices

The United States National Academies of Sciences, Engineering, and Medicine published a report in 2018 (United States National Academies of Sciences, Engineering, and Medicine 2018) to introduce the concept of *Open Science by Design*, in which a series of improvements were recommended to be implemented throughout the entire research life cycle to ensure open science and open data. To facilitate data sharing and to promote open science, some initiates listed below were recommended:

Data Generation

During data generation, researchers should consider collecting data in a digital form other than noting down data on a paper copy, e.g., a laboratory notebook. Many researchers are now collecting data in electronic forms (e.g., comma-separated values or CSV files). In addition, researchers should use tools compatible with open data, and also adopt automated workflows to format and curate generated data. These actions taken at the early stage of research life cycle can help avoid many problems in data sharing later on.

Data Sharing

After finishing preparing data, researchers should pick one or more data repositories to share their data. Data to be shared include not only data but also metadata and more. For example, World Data System (2015) recommended that data, metadata, products, and information produced from research all should be shared although national or international jurisdictional laws and policies might apply. Researchers should consult with funders or publishers about recommended data repositories into which they can deposit data. One example list of widely used data repositories (both general and discipline-specific) can be found here: https://www.nature.com/sdata/policies/repositories.

Conclusion

Data sharing act as a bridge linking both data providers and users, and it is particularly encouraged in the research community. Data sharing can benefit the research community in many ways including (1) improving the reproducibility and visibility of research, (2) promoting collaboration, and inspiring new science ideas, and (3) shared data can be used as a vehicle to foster the communication between academia, industry, and the general public (e.g., Brantley et al. 2018). To facilitate effective data sharing, researchers should follow FAIR principles (findability, accessibility, interoperability, and reusability) when they generate, format, curate, and share data.

Cross-References

▶ Data Repository

Further Readings

Brantley, S. L., Vidic, R. D., Brasier, K., Yoxtheimer, D., Pollak, J., Wilderman, C., & Wen, T. (2018). Engaging over data on fracking and water quality. *Science, 359* (6374), 395–397.

Popkin, G. (2019). Data sharing and how it can benefit your scientific career. *Nature, 569*(7756), 445.

United States National Academies of Sciences, Engineering, and Medicine. (2018). Open science by design: Realizing a vision for twenty-first century research. National Academies Press.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data, 3*, 160018.

World Data System. (2015). World Data System (WDS) Data Sharing Principles. Retrieved 22 Aug 2019, from https://www.icsu-wds.org/services/data-sharing-principles.