Open Research Data

Sarah Callaghan

British Atmospheric Data Centre
sarah.callaghan@stfc.ac.uk

Abstract. Open Research Data - A step by step guide through the research data lifecycle, data set creation, big data vs long-tail, metadata, data centres/data repositories, open access for data, data sharing, data citation and publication.

Keywords: Open Research Data, Open Access.

Lecture notes

The principles of Open Access are equally (if not more) applicable to research data. If the conclusions proposed by a research article are to be evaluated, not only must the article be made available, but also the data and methodologies (e.g. software, workflows, experiment descriptions) underlying those conclusions should be accessible for scrutiny and verification.

For data to be made available, infrastructure and services need to be put into place to support this. Anyone can upload data files to a website, or cloud-based storage. The difficulty lies in curating and managing the data, ensuring that it will remain available and accessible for the long term.

Data have many issues associated with their management that literature does not. For a start, journal articles are designed to be read by humans, not interpreted by machines, whereas for data the opposite is usually true. For a computer to be able to open and work with a set of given data files, there is a significant amount of metadata (data about data) that needs to be understood, for example: the format of the files (ASCII? Binary? Other formats – obsolete or otherwise?), the structure of the content within the files or database, and what the contents mean (variable names, calibrations, etc.) Datasets are often dynamic, meaning they have additions or changes made to them over time, some with greater version control than others. Researchers will work with old versions of a dataset because they don't want to wait until the final complete version is released, because that may happen years into the future.

Data is also far more heterogeneous than literature. Datasets may vary in size from single measurements to multiple TB (or more). Data may be timeseries, maps, the results of large computer models, images, recordings, or even physical objects. Data may be created by a single researcher working in isolation, or large teams, or even entire organisations. Data comes from all fields of research, from astronomy to zoology. All of this variability means that there is no "one size fits all" solution that can be applied to data.

Research funders have come to understand the importance of data, and put into place policies and fund infrastructure to encourage researchers to make their data open, thereby increasing data use and re-use, and improving the potential for collaboration. Building these infrastructures and policies has required significant collaboration between all of the parties interested in data – not only the researchers and funders, but also the institutional and discipline-specific data repositories, librarians, and academic publishers.

Part of the drive to make data more open acknowledges the extra work that a data producer must do to make their data open. Data citation and data publication are proposed as mechanisms to provide credit to the researcher for opening their data, and providing the metadata needed to ensure that the data can be used by other researchers.

Open access data, like open access literature, requires a cultural change in the way we conduct and manage research, but this is a change that will improve the research record for the better by allowing scientific reproducibility, and providing increased support for data reuse and cross-domain collaboration.