https://doi.org/10.1038/s41592-023-01865-4

Data sharing is the future

We examine data sharing practices and explore possible future directions.

In late 2022, the US government mandated open-access publication of scholarly research and free and immediate sharing of data underlying those publications for federally funded research beginning no later than 2025. For some fields the necessary standards and infrastructure are largely in place to support these policies. For others, however, many questions remain as to how these mandates can best be met.

In this issue, we feature a Correspondence from Richard Sever that was inspired by the government mandate and the increasing demand for open science. In it, he raises important topics, including deciding which data must be shared, standardizing file formats and developing community guidelines. He also calls for a "federated system of repositories with functionality tailored to the information that they archive," to meet the needs of many distinct fields.

Nature Portfolio journals have several data deposition requirements. These largely cover fields where data sharing has been standard practice for years. We strongly support data sharing and expect our authors to make data available immediately upon publication as well as over the long term after publication. We also actively ask authors to avoid 'data available upon request' statements except for exceptionally large datasets. Our papers also have stand-alone data availability statements to help to guide readers to source data. In addition, we are beginning a new collaboration with FigShare to host larger source data associated with our papers beginning at the peer review stage.

The fields of genomics and transcriptomics probably come closest to representing a model for data sharing, as consensus guidelines exist regarding types of data that must be shared and the form in which those data should be stored. Appropriate repositories are available for DNA and RNA sequences, genetic variation data, functional genomics data and gene expression data. The history of data sharing in genomics makes data storage and sharing the expectation from the onset of experimental design. A caveat is that genomics is expanding rapidly, especially with the rapid rise of spatial omics technologies, which have their own unique requirements for sharing and for which relatively few databases exist. As these methods emerge and grow, and as they become increasingly multimodal, new standards and databases may be needed.

Proteomics and structural biology are comparably mature. There are established repositories for protein sequences and proteomics data, and structural databases associated with crystallography, nuclear magnetic resonance structure determination and cryo-electron microscopy. And again, in these fields, data sharing formats are largely agreed upon, and sharing is often mandated by publishers and is certainly expected among researchers.

Other fields are still in the process of developing best practices for data sharing. For example, immunology research involves diverse methodological approaches and data types. As such, there is no one 'catch-all' repository for immunological data, nor are there many mandatory data sharing requirements apart from those for omics data. That being said, repositories are available that cover many widely used data types, such as flow cytometry data, immunogenomics and immune receptor repertoires. Efforts in this field are underway to further develop and implement best-practice guidelines for data sharing and also to improve the diversity and quality of data in databases.

Neuroscience is another field with diverse data storage and sharing needs, where a single common repository for all neuroscience data may be difficult to envisage given the differing needs of, for example, magnetic resonance imaging, microscopy, behavioral and electrophysiology data. Nevertheless, there has been substantial progress in the development of high quality, reliable databases and a strong community effort to promote data sharing. For example, the International Neuroinformatics Coordinating Facility has developed an infrastructure portfolio to help researchers to find solutions for their data sharing needs, including structural and functional neuroscience data from multiple modalities, large-scale projects and neurogenomics data.

Microscopy does not have a long history of data sharing, and most journals have no microscopy data deposition mandates. The challenges this field faces are many and include, huge dataset sizes, diverse data output from different modalities, questions surrounding what counts as 'raw data', the need to store and save multiple versions of files due to data processing, optimal file formats, best practices for metadata recording, and cost. However, groups like Quarep-LIMI, REMBI, Global BioImaging, Bioimaging North America and more are developing guidelines for data reporting and sharing that, should enable meaningful sharing and reuse of bioimaging data. And although not yet meeting the needs of all microscopists, image data resources and repositories such as the Image Data Resource and Bioimage Archive are growing and setting standards for the field.

A few themes emerge when examining data storage and sharing solutions across different fields. For one, data size matters a great deal to the feasibility of long-term data storage, let alone data sharing. Resources such as FigShare and Zenodo are becoming widely used, but they may have associated costs, especially for very large datasets. Moreover, issues of data privacy are paramount for many types of data involving human subjects and must be considered a top priority. Issues of data provenance and metadata standards are also crucial when it comes to reuse of data. Expectations within a field for data sharing are important for experimental planning, to create data that are not only shared but are also actually reusable. Perhaps the clearest theme of all, however, is that fields that share data as a matter of routine are richer for it, especially in the age of big data and artificial intelligence. Data sharing and reuse are more important now than ever.

We think the best path forward for all researchers involves smart guidelines and community consensus best practices to avoid ad hoc data storage and sharing solutions and promote reproducibility and reuse of data of all types. We hope that grant-funding agencies take note of the great needs of these diverse communities and continue to fund and develop stable databases to help to take storage and sharing burdens off the shoulders of individual laboratories.

Published online: 12 April 2023