Strategies for preserving and sharing data in biological sciences

THE UNIVERSITY of EDINBURGH
Centre for Engineering Biology

BioRDM
https://www.ed.ac.uk/biology/research/facilities/research-data-management

Livia C. T. Scorza, Tomasz Zieliński, Andrew J. Millar

Centre for Engineering Biology and School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JF, UK.

BioRDM: Biological Research Data Management team

Managing and sharing research data is a demanding task. The main barrier for sharing data routinely and in a reusable form are:

- The lack of expertise in Open and FAIR data principles;
- The lack of standards practices in data organization, storage and archiving;
- The lack of tools for seamless data description and handling;
- The diversity of experiments and data types in biology.

Our team assist researchers with their data needs by providing training and consultations, data curation services and creation or integration of software tools.
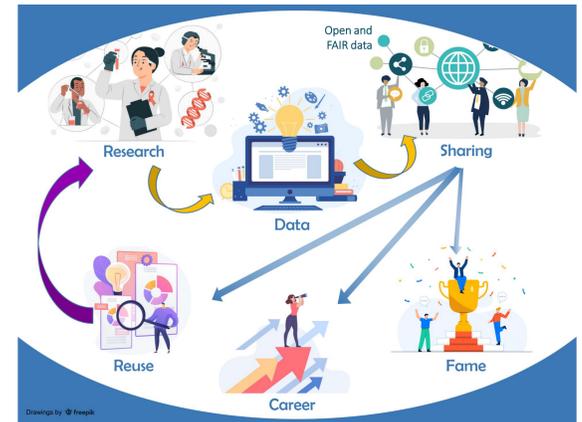
Our main mission is to promote the use of **Open Science** and **FAIR data principles** (Findable, Accessible, Interoperable and Reusable), in daily research workflows.

Data curation: selection, transformation and organization of data

The data curation process involves different steps, such as:

1. Select the research outputs for sharing (e.g., images, quantitative data, protocols, code, scripts, simulations);
2. Make data complete, unambiguous and described with a sufficient level of detail (rich metadata);
3. Export data to open formats (e.g., CSV, TIF, FASTA) when appropriate;
4. Make data easily findable by following naming conventions and organizing data in a sensible folder structure.

Sharing data:

After data curation, datasets can be shared in online repositories such as:

- General data agnostic repositories such as **Zenodo** (Fig.1), where different types of data (including posters!) can be uploaded;
- Data specific repositories (e.g., GenBank, BioImage archive);
- **protocols.io**: a platform for sharing protocols with a DOI, while tracking provenance of protocol versions (Fig.2);

- **GitHub:** Platform for sharing both code and data. It supports GitHub pages for creation of **project websites** enriching user's experience. (Fig. 4).

Full datasets can also be published in **open access journals dedicated to data**, such as *Scientific Data, Nature* (Fig. 3).
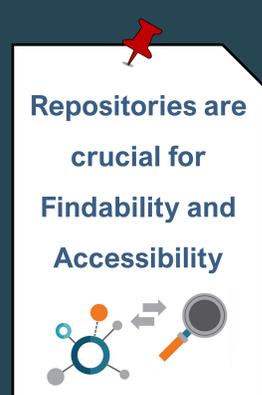
Repositories are crucial for Findability and Accessibility
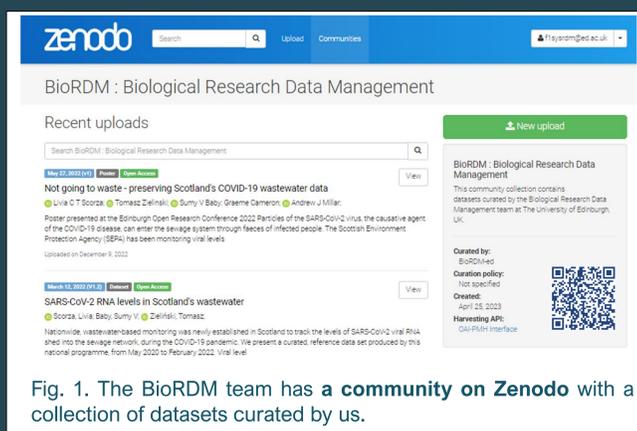
Fig. 1. The BioRDM team has **a community on Zenodo** with a collection of datasets curated by us.
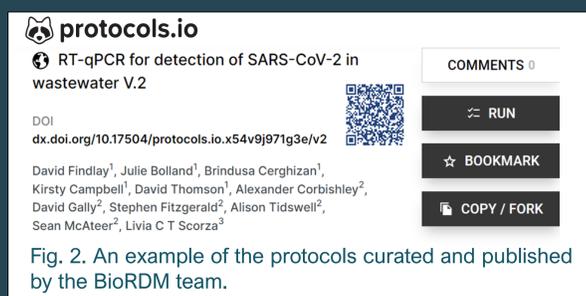
Fig. 2. An example of the protocols curated and published by the BioRDM team.
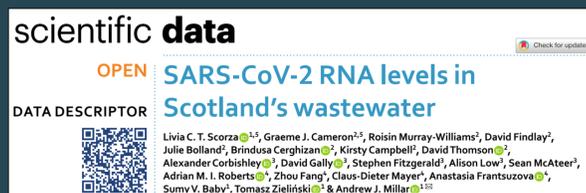
Fig. 3. A complete dataset as a peer-reviewed publication
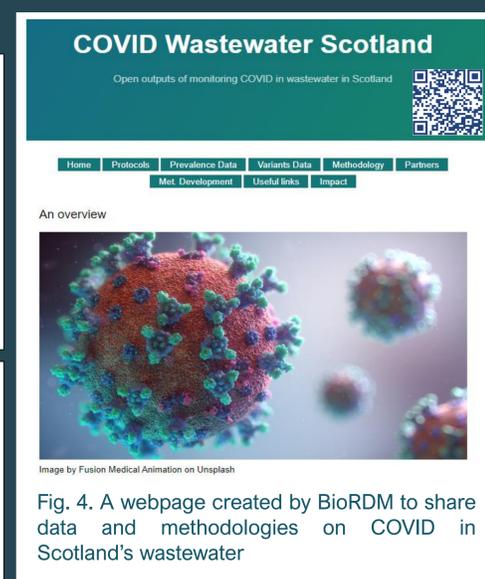
Fig. 4. A webpage created by BioRDM to share data and methodologies on COVID in Scotland's wastewater

FAIR in (biological) practice
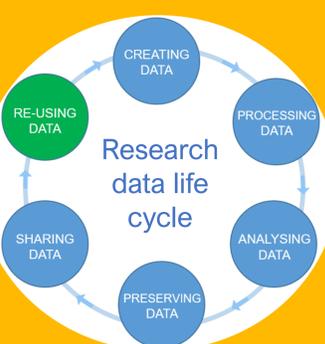
A hands-on workshop (also reused/ adapted by other institutions worldwide) helping researchers to practically apply Open Science and FAIR principles in their projects.

Typically, participants claim that they will incorporate new actions in their work such as: Improve their data description (metadata); Plan on how their folders/file's structure; Standardize file names; Write data management plans for all projects; Use Git version control system.

Increasing outputs

Sharing data as above will:

- ✓ Create extra research outputs for your CVs, as repositories emit DOIs for each dataset with their own metrics;
- ✓ Enhance your scientific contributions and achievements.