Supplementary material

Data Management Tools

Using general-purpose open repositories operating under the FAIR principles for non-sensitive project outcomes (such as research papers, data sets, research software, reports, slides, posters) can improve the overall project data management. An example is the Zenodo multi-disciplinary open repository maintained by CERN. A digital object identifier (DOI) is automatically assigned to all files (and new versions of files) uploaded in Zenodo.

The Research Data Management toolkit for Life Sciences, RDMkit, developed by ELIXIR with contributions from other European Biomedical Research Infrastructures, is a community-lead research data management toolkit for best practices and guidelines to support FAIR policies in data management. The RDMkit provides narrative context and explanations for steps in the data lifecycle, including data management planning; common data tasks, such as data brokering; specialised data management for domains and data types, such as bioimaging and rare diseases; and examples of RDM tool assemblies that support data journeys. The RDMkit provides a signposting gateway to resources such as the FAIRCookbook and the DSW common knowledge model, as well as smart contextual indexing into registries for training materials (the TeSS training portal); tools (Bio.tools); standards (FAIRsharing) and workflows (WorkflowHub). The EOSC4Cancer project, in a close collaboration with the CanServ project will produce a dedicated 'Cancer Data view' in the RDMkit that captures good practices, training resources and connects to data experts across Europe.

The Infectious Diseases Toolkit (IDTk), developed in the context of the BY-COVID project, provides a convergence point for knowledge exchange on efforts made in research response to infectious diseases. The IDTk aims at collecting and showcasing past and present responses to challenges on data handling, analysis and visualisations, while extracting best practices and guidelines to enhance preparedness. In addition, national resources and experiences can be showcased and shared even when direct access cannot be provided. The IDTk adopts the same ways of working and technical background as the RDMkit. As such, the IDTk is able to connect users with multiple tools and resources, including the RDMkit itself. The IDTk is a key asset open for the ISIDORe partners and users to showcase their data journeys, identify common solutions, and direct to the appropriate resources.

The FAIR Cookbook (Rocca-Serra et al. 2022) is an online, open and live resource with recipes that help users to make and keep data FAIR. Developed by data professionals and FAIR experts from ELIXIR, as well as the academic and industry sectors, the FAIR Cookbook guides users through the key steps of a FAIRification journey via recipes, which provide levels and indicators of FAIRness, the maturity model, the technologies, the tools and the standards available, as well as the skills required, and the challenges, to achieve and improve FAIRness. Each recipe tells you the audience type, reading time, level of difficulty, and the level of FAIR maturity it allows you to reach. Recipes are citable via their unique identifier, and their authors are credited, and have cross-references to FAIRsharing (for standards and repositories), RDMkit (for additional reading material), the DSW (to ensure FAIRness by design) and other resources in the ELIXIR ecosystem, and beyond. Supported by ELIXIR Nodes, and strong participation from the NIH Office of Data Science Strategy and major large pharmaceutical companies, the FAIR Cookbook is open to contributions by the ISIDORe community, which can write recipes to showcase community tools and resources, as well as examples of FAIRified data, to help with training, and sharing of common practices.

The EOSC service WorkflowHub is a registry of computational workflows, developed by projects including EOSC-Life and BY-COVID that initially targeted Life Sciences and COVID-19 workflows (Goble et al. 2021). The workflows and computational notebooks deposited in the hub now cover a wide range of communities including biodiversity and earth sciences. Registries are essential to support FAIR computational workflows (Goble et al. 2022), such as persistent identifiers, versioning, attribution and detailed metadata. As many ISIDORe data practices involve computational analysis, capturing the implied workflow as a method is important for reproducibility and reuse.

RO-Crate is a community-developed method for lightweight packaging of diverse scientific research data with structured metadata, based on FAIR principles and Linked Data standards JSON-LD and schema.org (Soiland-Reyes et al. 2022a). RO-Crate has been integrated with machine-actionable Data Management Plans by using DMPs as templates for building FAIR data packages (Miksa, Jaoua & Arfaoui 2020), and using exemplar crates to kick-start populating DMPs (Soiland-Reyes et al. 2022b), both approaches now being adopted by ELIXIR’s DSW (Eguinoa et al. 2022). RO-Crate has rich support for describing software and workflows, utilised by the ELIXIR Software Management Plans (Giraldo et al. 2022) and EOSC-Life workflow ecosystem including WorkflowHub (Goble et al. 2021). ISIDORe can utilise RO-Crate as a way to connect DMPs and data deposits with extensible metadata profiles for different domains, capturing lightweight data derivation provenance with the Common Provenance Model (CPM) (Wittner et al. 2022).

The CPM is a provenance model developed under the auspices of the EOSC-Life project. The main purpose is to support the traceability of data precursors in distributed environments, address reproducibility issues, and enable a more effective data quality and fit-for-purpose assessment. The model is designed for complex, heterogeneous, and multi-institutional environments with different data access restrictions for sensitive data, and subsequently, their provenance. CPM solves the access and reputability challenge by creating multiple provenance bundles, each documenting one environment data and its precursors, such as biological material or environmental specimens, and links these bundles into a common provenance chain. In addition, the model forms a conceptual foundation for an ISO 23494 series (Wittner et al. 2021) (still under development) and is being integrated with the RO-Crate specification within the BY-COVID project (Soiland-Reyes et al. 2022a). Finally, the standard and the underlying model aim for a very high level of provenance interoperability within life sciences domains. Despite its simplicity, usage of the model requires some decisions and design considerations, which may be described in a DMP.

FAIRsharing is a community-driven, cross-disciplinary resource, with users and collaborators across all disciplines, that describes community-driven standards, databases, repositories and data policies and the relationships among them. It is also a recommended interoperability resource within ELIXIR and EOSC-Life, and an output of the RDA FAIRsharing WG. Plans are to work with ISIDORe to create a branded FAIRsharing collection that displays live descriptions and network graphs of the resources recommended within their DMP as many have already done, such as IVOA, FAIRsharing EOSC-Life and the ISO 20691 specification. Next, we will also ensure that key ISIDORe representatives join the FAIRsharing community curation programme to ensure that resources developed and used by ISIDORe are described as richly as possible, providing the ISIDORe community with a well-curated landscape graph of resources relevant to them. Traversal of the ISIDORe graph can aid resource discovery, gap analysis and further targeted outreach and engagement. Additionally, any data management policies created by ISIDORe will be added to FAIRsharing to increase the findability, accessibility and reusability of the policy descriptions through the FAIR data policy 'workflow' described earlier.

Definitions

This section aims to provide definitions of common terms used in the document (the source is Kesisoglou et al. 2022, unless otherwise stated).

Data management plan (DMP): A DMP is a structured document that describes data management during and after completion of a research project. It includes information on data capacity, data production, data quality, data safety and protection measures, as well as a role description of people dealing with these tasks. Through recommendations and curation tools, DMPs make data available for reuse in a sustainable way (Bishop et al. 2020). Good research data management prevents data loss, ensures description for appropriate reuse through metadata, keeps data secure and facilitates data sharing. National and international entities have put forward specific requirements to be included in the DMP. In August of 2022, the White House Office of Science and Technology Policy (OSTP) announced that, by 2025, scientific data from all new federally funded research must be made accessible to the US public. Similarly, the European Commission pushes for open access to scientific information, including support for the development of the European Open Science Cloud (EOSC) with major investment from the European Horizon 2020 and following Horizon Europe research and innovation programmes.

Data policy: A set of broad, high-level principles which form the guiding framework in which data management can operate.

Dataset: Collection of data that is represented in a particular form. Datasets will vary depending upon the type of intended use, and how the collecting organisation has decided to organise their data upon collection. Dataset is essentially a heterogeneous term that could be made up of any type of collection for any type of data.

Data steward: A person who has an administrative role; they do not really use the data. They create guidelines to make data FAIR and advice on how to do it. Stewards might have direct responsibility on the data at hand (processors) or not.

Metadata: A set of data that defines and describes a resource (e.g., data, dataset, sample) so that it can be understood, discovered and reused. There are different levels of metadata. Since metadata can be used to describe different aspects of data, we can group metadata properties in terms of quality, availability, provenance, processing, among others. Then there are metadata catalogues that can be developed to describe the available data collections in a repository or hub. Metadata is important to make data understandable, and can contribute to increase the findability, accessibility, interoperability and reusability of the data. Metadata can be collected or compiled in repositories to improve the FAIRness level of the data collections.

Personnel data: The stated aim of the GDPR was to enable data subjects to have greater control over their 'personal data', whilst unifying the European data protection rules. The definition of 'personal data', as outlined in Article 4(1) of GDPR, includes 'any information relating to an identified or identifiable natural person (data subject)'. This includes names, surnames, home address, email address, or an identifier number or data held by a hospital or laboratory that could be used to identify a living individual. In addition, the existence of special categories of personal data, referred to as sensitive personal data, adds another layer of complexity. Sensitive personal data are outlined in Article 9(1) GDPR and include data pertaining to ethnicity, sexual orientation, religious beliefs, trade union membership, and genetic data. Genetic data is defined as 'personal data relating to the inherited or acquired genetic characteristics of a natural person which result from the analysis of a biological sample from the natural person in question, in particular chromosomal, deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) analysis, or from the analysis of another element enabling equivalent information to be obtained'.

Quality assurance and control: Regardless of whether a project re-uses / creates new data, or not, their quality assurance and control throughout and beyond the project must be reflected in the DMP. This can be particularly challenging in those consortia with a large number of participants.

Quality ultimately depends on the purpose, so the data flow must be clearly stated in the DMP, indicating the specific objectives of the different working groups and the data use/management in each of them. Quality measures related to each of these steps have to be taken into account in the DMP, without losing the focus on the overarching structure and main goals of the entire project. In addition, not only the quality control procedures must be reflected in the DMP, but also where they are going to be collected. This is important in order to ensure that the data as well as the QA/QC procedures follow the FAIR principles and are available for other projects (discussion from ISIDORe).

Research infrastructures: The European Commission (EC) defines research infrastructures (RIs) as facilities that provide resources and services for research communities to conduct research

and foster innovation. They can be used beyond research, for example, for education or public

services, and they may be single-sited, distributed, or virtual. They often include: i)

major scientific equipment or sets of instruments, ii) collections, archives or scientific

data, iii) computing systems and communication networks, and iv) any other research and

innovation infrastructure of a unique nature that is open to external users.

Sensitive data: Information that is regulated by law due to possible risk for plants, animals, individuals and/or communities and for public and private organisations. Sensitive personal data include information related to racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership and data concerning the health or sex life of an individual. These data could be identifiable and potentially cause harm through their disclosure.