A FAIRification roadmap for ELIXIR Software Management Plans

Olga Giraldo[‡], Renato Alves[§], Dimitrios Bampalikis[|], Jose M Fernandez[¶], Eva Martin del Pico[¶], Fotis E Psomopoulos[#], Allegra Via[¤], Leyla Jael Castro[‡]

‡ ZB MED Information Centre for Life Sciences, Cologne, Germany
§ European Molecular Biology Laboratory, Heidelberg, Germany
| National Bioinformatics Infrastructure Sweden, Uppsala, Sweden
¶ Barcelona Supercomputing Center, Barcelona, Spain
# Centre for Research and Technology Hellas, Thessaloniki, Greece
¤ Institute of Molecular Biology and Pathology - National Research Council, Rome, Italy

Corresponding author: Leyla Jael Castro (ljgarcia@zbmed.de)

Abstract

Academic research requires careful handling of data plus any means to collect, transform and publish it, activities commonly supported by research software (from scripts to end-user applications). Data Management Plans (DMPs) are nowadays commonly requested by funders as part of good research practices. A DMP describes the data management lifecycle for the data corresponding to a research project, covering activities from collection to publication and preservation. To support and improve transparency, open science, reproducibility (and other *ilities), data needs to be accompanied by the software transforming it. Similar to DMPs, Software Management Plans (SMPs) can help formalize a set of structures and goals ensuring that the software is accessible and reusable in the short, medium and long term. DMPs and SMPs can be presented as text-based documents, guided by a set of questions corresponding to key points related to the lifecycle of either data or software.

A step forward for DMPs are the machine-actionable DMPs (maDMPs) proposed by the Research Data Alliance DMP Common Standards Working Group. A maDMP corresponds to a structured representation of the most common elements present in a DMP (Miksa et al. 2020b), overcoming some obstacles linked to text-based representation. Such a structured representation makes it easier for DMPs to become readable and reusable for both humans and machines alike. The DMP Common Standard ontology (DCSO) (Cardoso et al. 2022) further supports maDMPs as it makes it easier to extend the original maDMP application profile to cover additional elements related to, for instance, SMPs or specific requirements from funders. maDMPs can be combined with the notion of a Research Object Crates (RO-Crate) to automate and ease management of research data (Miksa et al. 2020a). An RO-Crate (Soiland-Reyes et al. 2022) is an open, community-driven, and lightweight approach based on schema.org (Guha et al. 2016) annotations in

JSON-LD to package research data (or any other research digital object) together with its metadata in a machine-readable manner.

The ELIXIR SMP has been developed by the ELIXIR Software Development Best Practices Group in the ELIXIR Tools Platform to support researchers in life sciences (Alves et al. 2021). The ELIXIR SMP aims at making it easier to follow research software good practices aligned to the findable, accessible, interoperable and reusable principles for research software (FAIR4RS) (Chue Hong et al. 2022) while dealing with the lifecycle of research software. Its primary goal is encouraging a wider adoption by life science researchers, and being as inclusive as possible to the various levels of technical expertise. Here we present a roadmap for ELIXIR SMPs to become a FAIR digital object (FDO) (Schultes and Wittenburg 2019) based on the extension of maDMPs and DCSO and the use of RO-Crates. FDOs have been proposed as a way to package digital objects together with their metadata, types, identifiers and operations, so they become more machine-actionable and auto-contained.

The current version of the ELIXIR SMP includes seven sections: accessibility and licensing, documentation, testing, interoperability, versioning, reproducibility, and recognition. Each section includes questions guiding and supporting researchers so they cover key aspects of the software lifecycle relevant to their own case. To lower the barrier and make it easier for researchers, most questions are Yes/No with some few offering a set of options. In some cases, a URL is also requested, for instance regarding the location of the documentation for end-users. Our roadmap for ELIXIR SMPs to move from a text-based questionnaire to an FDO comprises four main steps:

1. creating maSMP application profile,
2. extending DCSO,
3. mapping to schema.org, and
4. using RO-Crates.

Our maSMP application profile will include the semantic representation of the structured metadata that comes from the ELIXIR SMP. We will add granularity to the current root of the DCSO (dcso:DMP), by proposing the term SMP. In addition, we will propose the term ResearchSoftware as a dcso:Dataset. Terminology related to documentation, such as “Objective" will also be considered. The objective is the *Why* the research software, which is crucial for their comprehensibility. We will propose the term DatasetObjective as the reason for the creation of a dataset. Source-codeRepository and Source-codeTesting are also good candidates to be part of the DCSO extension.

We will extend DCSO with new classes and properties as necessary to include the software related elements mentioned in the maSMP application profile. As the ELIXIR SMP targets the life science community, we will analyze the need to add links from DCSO to ontologies describing common operations, activities, and types in this domain. One important aspect is the creation of a mapping from DCSO to schema.org. Schema.org has become a popular choice to add lightway semantics to web pages but can also be used on its own to provide metadata describing all sorts of objects. In life sciences, Bioschema

s (Gray et al. 2017) offers guidelines on how to use some of the schema.org types aligned to this domain. Bioschemas includes a set of profiles, including minimum, recommended and optional properties, that have been agreed to and adopted by the community, for instance the ComputationalTool profile provides a way to describe software tools and applications. Bioschemas promotes its adoption by key resources in Life Sciences and development of tools such as the Bioschemas Markup Scraper and Extractor (BMUSE) used for the harvesting of the data (Gray et al. 2022).

Our final step for ELIXIR SMPs to become an FDO is using RO-Crates to package research software together with its metadata and link it to/from its corresponding SMP. To do so, we will create an RO-Crate profile capturing the metadata needed to describe software tools including elements from the SMP. It will become a versioned living crate as research software evolves with time, particularly when new releases are published. Thanks to the RO-Crate bundling nature, where digital objects are packed together with its metadata, a software crate enriched with the elements from the SMP are a good example of an FDO as all the critical information about a software tool is bound together in a unit that can be shared with peers via FAIR registries and repositories.

Keywords

Software Management Plans, RO-Crates, Research Software, FAIR4RS

Presenting author

Leyla Jael Castro

Presented at

First International Conference on FAIR Digital Objects, poster

Conflicts of interest

References

- Alves R, Bampalikis D, Castro LJ, González JMF, Harrow J, Kuzak M, Martin E, Psomopoulos F, Via A (2021) ELIXIR Software Management Plan for Life Sciences. BioHackrXiv. https://doi.org/10.37044/osf.io/k8znb
- Cardoso J, Castro LJ, Ekaputra F, Jacquemot M, Suchánek M, Miksa T, Borbinha J (2022) DCSO: Towards an Ontology for Machine-actionable Data Management Plans. DOI: 10.21203/rs.3.rs-1458035/v1. URL: https://www.researchsquare.com

- Chue Hong N, Katz D, Barker M, Lamprecht A, Martinez C, Psomopoulos F, Harrow J, Castro LJ, Gruenpeter M, Martinez PA, Honeyman T, Struck A, Lee A, Loewe A, van Werkhoven B, Jones C, Garijo D, Plomp E, Genova F, Shanahan H, Leng J, Hellström M, Sandström M, Sinha M, Kuzak M, Herterich P, Zhang Q, Islam S, Sansone S, Pollard T, Atmojo UD, Williams A, Czerniak A, Niehues A, Fouilloux AC, Desinghu B, Goble C, Richard C, Gray C, Erdmann C, Nüst D, Tartarini D, Ranguelova E, Anzt H, Todorov I, McNally J, Moldon J, Burnett J, Garrido-Sánchez J, Belhajjame K, Sesink L, Hwang L, Tovani-Palone MR, Wilkinson M, Servillat M, Liffers M, Fox M, Miljković N, Lynch N, Martinez Lavanchy P, Gesing S, Stevens S, Martinez Cuesta S, Peroni S, Soiland-Reyes S, Bakker T, Rabemanantsoa T, Sochat V, Yehudi Y, WG RF (2022) FAIR Principles for Research Software (FAIR4RS Principles). https://doi.org/10.15497/RDA00068
- Gray AG, Papadopoulos P, Gaignard A, Rosnet T, Mičetić I, Moretti S (2022) Bioschemas data harvesting project report. BioHackrXiv. https://doi.org/10.37044/osf.io/y6gbq
- Gray AJG, Goble C, Jimenez RC (2017) From Potato Salad to Protein Annotation. ISWC Posters and Demo session. URL: http://ceur-ws.org/Vol-1963/paper579.pdf
- Guha RV, Brickley D, Macbeth S (2016) Schema.org: evolution of structured data on the web. Communications of the ACM 59 (2): 44-51. https://doi.org/10.1145/2844544
- Miksa T, Jaoua M, Arfaoui G (2020a) Research Object Crates and Machine-actionable Data Management Plans. 1st Workshop on Research Data Management for Linked Open Science. https://doi.org/10.4126/FRL01-006423291
- Miksa T, Walk P, Neish P (2020b) RDA DMP Common Standard for Machine-actionable Data Management Plans. https://doi.org/10.15497/rda00039
- Schultes E, Wittenburg P (2019) FAIR Principles and Digital Objects: Accelerating Convergence on a Data Infrastructure. Communications in Computer and Information Science3-16. https://doi.org/10.1007/978-3-030-23584-0_1
- Soiland-Reyes S, Sefton P, Crosas M, Castro LJ, Coppens F, Fernández J, Garijo D, Grüning B, La Rosa M, Leo S, Ó Carragáin E, Portier M, Trisovic A, RO-Crate Community, Groth P, Goble C (2022) Packaging research artefacts with RO-Crate. Data Science1-42. https://doi.org/10.3233/DS-210053