1st Conference on Research Data Infrastructure

Enabling RDM

https://doi.org/10.52825/CoRDI.v1i.251

Published: 07 Sept. 2023

RDMkit: The Research Data Management Toolkit for Life Sciences

Nazeefa Fatima[1][https://orcid.org/0000-0001-7791-4984] , Pinar Alper[2][https://orcid.org/0000-0002-2224-0780], Federico Bianchini[1][https://orcid.org/0000-0002-9016-4820], Korbinian Bösl[3][https://orcid.org/0000-0003-0498-4273], Ulrike Wittig[4][https://orcid.org/0000-0002-9077-5664], Carole Goble[5][https://orcid.org/0000-0003-1219-2137], Frederik Coppens[6][https://orcid.org/0000-0001-6565-5145]

[1] University of Oslo, Norway

[2] Luxembourg National Data Service, Luxembourg

[3] University of Bergen, Norway

[4] Heidelberg Institute for Theoretical Studies, Germany

[5] University of Manchester, United Kingdom

[6] The Flemish Institute for Biotechnology, Belgium

Abstract. Effective data management extends over the entire life cycle of data, from the point of creation through to dissemination and archiving, and usually continues long after a research project is concluded. Tailored guidance for management of research data is increasing its importance among data-driven scientific investigations. There is now increasing demand for Research Data Management (RDM), by funders and host institutions, for researchers to develop and implement data management plans for projects. The RDMkit, by ELIXIR Europe, is an open community-led resource designed to share knowledge on how to carry out RDM in life sciences research. It fills the training and learning gap by providing knowledge on domain-focused RDM considerations, task-focused solutions and resources, and tool assemblies that showcase real-life examples addressing how combinations of tools can be used to go through the data life cycle. The technical infrastructure of the RDMkit is reproducible, making it possible for other organisations to use the structure as a guide and inspiration to set up RDM source for their users. In this talk, RDMkit will be introduced with example best practice guidelines, and integrations and knowledge sources.

Keywords: Crowdsourced, RDM, Knowledge Resource

1. RDMkit Overview

Research data management (RDM) is crucial for the reproducibility of studies and the re-usability of scientific results. RDM is based on a life-cycle view of data that extends beyond the timespan of a research study, preceding data's creation and often lasting beyond data and results dissemination. In order to foster RDM and embed it into the scientific process, funders require researchers to systematically plan for data management with Data Management Plans (DMPs).

Funders, research institutions and infrastructures, offer a multitude of support to help researchers in building and executing DMPs. The support landscape includes institutional data stewards who consult and train researchers, as well as RDM policies, guidelines, standards

and software tools from various sources. Two aspects of current support prevent it from reaching its full potential. First, for researchers, navigating this populous landscape, even within a single research infrastructure, can be overwhelming. Data stewards play a role here as advisors, however, they need mechanisms to effectively capture and disseminate their know-how. Secondly, most guidance found in current resources is often generic focusing on funder DMP templates, whereas surveys reveal that researchers need tailored, discipline-specific guidance and examples [1].

To address these challenges and bridge the knowledge gaps, ELIXIR (the pan-European research infrastructure for life-science data [2]) has convened a transnational community of bioscientists to build the RDMkit [3]. Launched in March 2021, the RDMkit is a collaborative effort by data stewards, life scientists, and RDM experts from all areas of biomedical sciences. It provides RDM best practice guidelines reachable through entry points based on the data lifecycle, the personas in RDM as well as various life science domains. Per topic, RDMkit highlights key considerations and the resources that can be used to build solutions in a way that makes life-science data Findable, Accessible, Interoperable and Reusable (FAIR). The RDMkit integrates with other ELIXIR tools and resources such as FAIRCookbook [4], FAIRsharing [5], bio.tools [6], Data Stewardship Wizard [7] and TeSS [8] training portal enabling the user to deliver a seamless data management plan and form an RDM knowledge commons for the ELIXIR community.

For researchers, the RDMkit is a one-stop open source of information, advice, and signposting to RDM know-how, tools, examples and best practices written by life scientists for life scientists. For data managers, RDMkit is a resource to complement institutional guidelines. For funding agencies and policymakers, RDMkit is a resource that can be included in guidelines. With over 100 webpages, RDMkit content spans a range of life science domains from metagenomics to human data and various tasks from metadata collection to data publication, considering all steps of the data management life cycle. In addition to generic advice, the RDMkit provides a collection of country-specific information resources such as local policies and national regulations on data ethics.

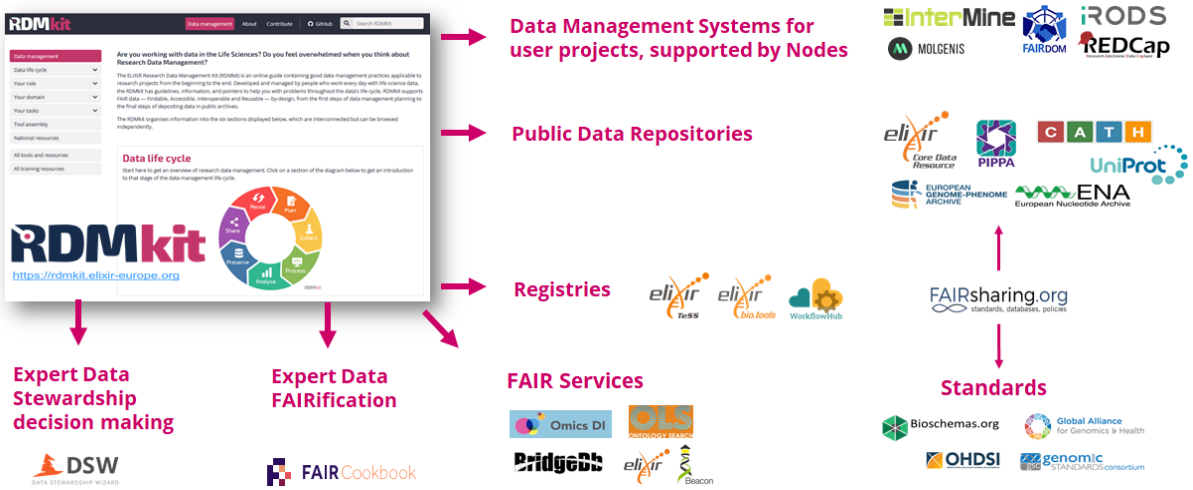

Figure 1. RDMkit offers exchange of FAIR know-how through integrations with other databases and resources, and signposting to resources in the community.

As a community-driven effort, coordinated by the Editorial Board, RDMkit is open to contributions from everyone. So far, the content is developed by over 150 contributors with references to over 340 tools and resources explained in the context of data management solutions to real-world problems. Other European Research Infrastructure providers - EuroBioImaging, BioExcel Centre of Excellence for Molecular Modelling - have contributed to specialist pages.

RDMkit has immediately gathered high visibility with 10K+ users from across 100 countries since its launch in March 2021 (Google Analytics), with users from across South America and the African continents as well as the USA and Europe. RDMkit is scalable and replicable, with open access code and content material available at GitHub. Although RDMkit pages are written in English, there is no limit to the languages for the content development.

Shortly after its release, the RDMkit has been recognised in funder guidelines in the EU. The European Commission's Horizon Europe Programme Guide recommends it as the "resource for Data Management guidelines and good practices for the Life Sciences", and it is listed in European Research Council guidelines for grantees. The RDMkit website is referred to by various national funders, universities and research institutes. Its open-source web infrastructure has been adopted by several national/international RDM initiatives, ranging from Australian BioCommons, to the 1+ Million Genomes Trust Framework, and the European BY-COVID project. The NIH Office of Data Strategy has representation on the RDMkit editorial board, for the possible alignment of NIH-funded RDM support activities with the RDMkit approach.

At the 1st Conference on Research Data Infrastructure (CoRDI), the RDMkit will be introduced with example best practice guidelines from its different sections. We will showcase the integration of RDMkit with other knowledge resources in ELIXIR. The talk will also describe RDMkit's contribution and editorial processes as well as its underlying technology and implementation.

Data availability statement

RDMkit content and site infrastructure is open-source and can be found at the following GitHub repositories

- ELIXIR Toolkit Theme: https://github.com/ELIXIR-Belgium/elixir-toolkit-theme
- RDMkit content: https://github.com/elixir-europe/rdmkit

Underlying and related material

Not applicable.

Author contributions

NF and PA wrote the original draft abstract.

CG and FC supervised the work and reviewed and edited the abstract.

Competing interests

The authors declare that they have no competing interests.

Funding

RDMkit has been developed as part of the Horizon 2020 Project ELIXIR CONVERGE Grant agreement ID: 871075.

The Luxembourg National Data Service (LNDS) is a brand of the Plateforme Nationale d'Échange de Données (PNED G.I.E), an economic interest group established by the Luxembourgish Government.

Acknowledgement

RDMkit's best practice guidelines have been developed in a crowdsourced fashion by an open community contributors listed on https://rdmkit.elixir-europe.org/contributors.

References

[1] Marjan Grootveld, Ellen Leenarts, Sarah Jones, Emilie Hermans, & Eliane Fankhauser. (2018). OpenAIRE and FAIR Data Expert Group survey about Horizon 2020 template for Data Management Plans (1.0.0). Zenodo. https://doi.org/10.5281/zenodo.1120245

[2] Harrow J, Drysdale R, Smith A, Repo S, Lanfear J, Blomberg N. ELIXIR: providing a sustainable infrastructure for life science data at European scale. Bioinformatics. 2021 Aug 25;37(16):2506-2511. doi: 10.1093/bioinformatics/btab481. PMID: 34175941; PMCID: PMC8388016.

[3] RDMkit Community. The Research Data Management toolkit for Life Sciences. https://rdmkit.elixir-europe.org/

[4] Philippe Rocca-Serra, Wei Gu, Vassilios Ioannidis, Tooba Abbassi Daloii, Salvador Capella-Gutierrez, Ishwar Chandramouliswaran, Andrea Splendiani, Tony Burdett, Robert T. Giessmann, David Henderson, Dominique Batista, Allyson Lister, Ibrahim Emam, Yojana Gadiya, Lucas Giovanni, Egon Willighagen, Chris Evelo, Alasdair J. G. Gray, Philip Gribbon, … the FAIR Cookbook Recipes' Authors. (2022). The FAIR Cookbook - the essential resource for and by FAIR doers (1.0). Zenodo. https://doi.org/10.5281/zenodo.7156792

[5] Sansone SA, McQuilton P, Rocca-Serra P, Gonzalez-Beltran A, Izzo M, Lister AL, Thurston M; FAIRsharing Community. FAIRsharing as a community approach to standards, repositories and policies. Nat Biotechnol. 2019 Apr;37(4):358-367. doi: 10.1038/s41587-019-0080-8. PMID: 30940948; PMCID: PMC6785156.

[6] Niall Beard, Finn Bacall, Aleksandra Nenadic, Milo Thurston, Carole A Goble, Susanna-Assunta Sansone, Teresa K Attwood, TeSS: a platform for discovering life-science training opportunities, *Bioinformatics*, Volume 36, Issue 10, May 2020, Pages 3290–3291, https://doi.org/10.1093/bioinformatics/btaa047

[7] Ison J, Ienasescu H, Chmura P, Rydza E, Ménager H, Kalaš M, Schwämmle V, Grüning B, Beard N, Lopez R, Duvaud S, Stockinger H, Persson B, Vařeková RS, Raček T, Vondrášek J, Peterson H, Salumets A, Jonassen I, Hooft R, Nyrönen T, Valencia A, Capella S, Gelpí J, Zambelli F, Savakis B, Leskošek B, Rapacki K, Blanchet C, Jimenez R, Oliveira A, Vriend G, Collin O, van Helden J, Løngreen P, Brunak S. The bio.tools registry of software tools and data resources for the life sciences. Genome Biol. 2019 Aug 12;20(1):164. doi: 10.1186/s13059-019-1772-6. PMID: 31405382; PMCID: PMC6691543.

[8] Pergl, R., Hooft, R., Suchánek, M., Knaisl, V. and Slifka, J., 2019. "Data Stewardship Wizard": A Tool Bringing Together Researchers, Data Stewards, and Data Experts around Data Management Planning. Data Science Journal, 18(1), p.59. DOI: http://doi.org/10.5334/dsj-2019-059