1st Conference on Research Data Infrastructure

Enabling RDM

https://doi.org/10.52825/CoRDI.v1i.421

Published: 07 Sept. 2023

Galaxy and RDM

Being more than a workflow manager: living the data life cycle

Sebastian Schaaf[1,2][https://orcid.org/0000-0003-2982-388X] , Anika Erxleben-Eggenhofer[1][https://orcid.org/0000-0002-7427-6478], and Björn Grüning[1][https://orcid.org/0000-0002-3079-6586]

[1] Freiburg Galaxy Team, Bioinformatics Group, Department of Computer Science, University of Freiburg, Germany

[2] ELIXIR Officers Team, Institute for Bio- and Geosciences 5, Research Center Jülich, Germany

Abstract. The increasing amount of data generated by scientific research poses the challenge of providing an adequate infrastructure and tools that facilitate FAIR (Findable, Accessible, Interoperable and Reusable) data access, manipulation, analysis and visualization. Often, the burden of managing the metadata associated with the original data and the analysis lies with the researchers.

The open source Galaxy platform [1] is well-known for supplying tools and workflows for reproducible and transparent data analysis across scientific disciplines. It is a multi-user environment which facilitates sharing of e.g. tools, workflows, notebooks, visualizations, and data with others. There are three large Galaxy instances (US, Europe [2] and Australia) used by hundreds of thousands of researchers worldwide and that are using PBs of data. Galaxy handles the metadata transparently, releasing scientists from the burden and making it less prone to human errors. These features can be used without technical background by using a web browser or, for experts, through the Galaxy API.

Galaxy is more than a workflow manager: it provides scientists with access to reference data, databases (ENA, UniProt, NCBI, PDB, Ensembl…), external repositories (FTP, SFTP, Dropbox…), data sources through standard APIs (TRS, DRS from GA4GH). Importantly, Galaxy not only uses the metadata of input data, but enriches the metadata space with analysis information. Thus, users are finally enabled to apply export mechanisms for the data, targeting external resources (S3, ENA…). Beyond, also the applied workflow invocation can be exported invoking standardized formats (RO-Crate, BioComputeObject …; [3]), effectively easing the sharing of the research artifacts, but also the details on the underlying analysis.

Although easy and standardized import/export functionalities are crucial features, entire data life cycles can be mapped into Galaxy. From the users' perspective, the built-in sharing features for data, workflows, histories etc. appear much more relevant in daily research efforts, making it particularly easy to reproduce results in order to verify their correctness and enable other researchers to build upon them in future studies. In fact, one of Galaxy's key features is its emphasis on **transparency, reproducibility,** and **reusability**. All provenance information of a dataset, including versions of used tools, parameters, execution environment, is captured and can be reused or exported using standards like BCO or RO-Crate to public archives. This highest level of provenance tracking also enables **traceability** and can also be used to reduce the environmental impact of a data analysis.

In addition to reproducibility, the Galaxy project places a strong emphasis on research data management (**RDM**; [4]). Beyond platform tools for data import, organization, sharing,

annotation, and export, data can be stored and accessed through a variety of providers, including cloud storages like NextCloud. This allows researchers to work with large datasets without the need for local storage infrastructures. The project encourages researchers to share their data and analysis workflows with the wider scientific community, with the aim of accelerating scientific discovery and innovation by following the FAIR principles. Notably, public instances offer not only computing capacities to users, but also persistent disk space, decoupling researchers from dependencies on local capacities and technical challenges (resources, capacities, support, …). This enables democratizing data analysis in large. In practice, centralized and efficient user support is an important aspect of intuitive and borderless sharing in data science. Failures in tools or analysis procedures can be fixed for numerous users simultaneously and users can be nudged to use more efficient or correct tools.

In terms of RDM in Galaxy we will put a focus on **RO-Crate** (Research Object Crate; [5]) as a relatively recent development, implementing to a practical extent the FAIR Digital Objects (FDO) concept. RO-Crate enables researchers to organize and package their research data and other digital resources in a way that makes it easier to share, reuse, and reproduce their work; obviously this shows a great overlap with Galaxy's principles. On the European level, ELIXIR mandates the increased usage of RO-Crate. Notably, also in the German NFDI space decisions have been made for pushing RO-Crate, with DataPLANT being an early adopter. A central mission of DataPLANT, the first-round NFDI-funded consortium for connecting plant researchers in Germany, is to provide a suitable infrastructure for data analysis.

Galaxy has been widely adopted by the research community, particularly in fields such as data science, bioinformatics, and digital humanities. It has also been recognized as a key pillar of various European organizations such as the European Open Science Cloud (EOSC). The Freiburg Galaxy Team, being a central pillar for operating the European Galaxy server 'UseGalaxy.eu', is part of these efforts and also involved in further NFDI consortia. The European Galaxy server is a flagship project of the German Network of Bioinformatics infrastructure (de.NBI) and part of multiple CRCs. Since 2022 the European Galaxy community is officially part of EOSC with its own project 'EuroScienceGateway' [6], where OpenAIRE, the EOSC organization behind Zenodo, is an important partner. They bring in their knowledge graph, aggregating research data properties (metadata, links), supporting open science principles.

Our talk will describe how the Galaxy platform assists researchers from diverse technical backgrounds and scientific domains throughout the entire data life cycle: data access, processing, analysis, preservation, sharing and re-use (Figure 1). We will highlight cross-connections with other popular services and sketch future directions.

Keywords: RDM, Galaxy, data life cycle, reproducibility, RO-crate

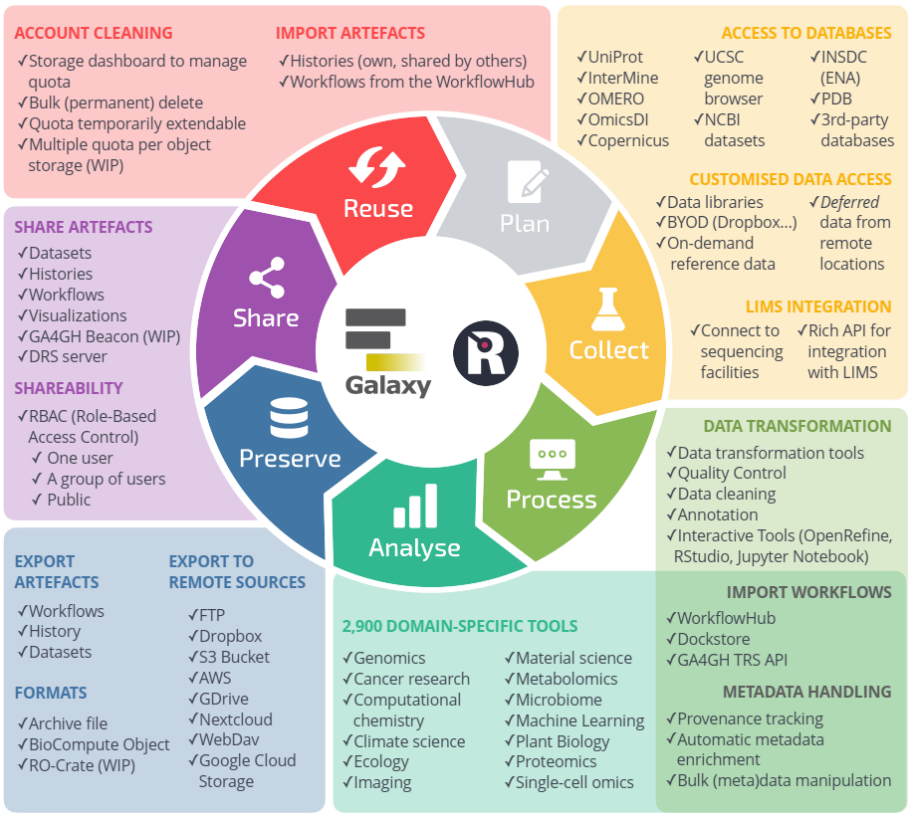

Figure 1. Data life cycle invoking Galaxy and RO-crate.

Competing interests

The authors declare that they have no competing interests.

References

1. The Galaxy Community - Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update; Nucleic Acids Research, Volume 50, Issue W1, 5 July 2022, Pages W345–W351, https://doi.org/10.1093/nar/gkac247
2. "Galaxy | Europe." https://usegalaxy.eu/ (accessed Apr. 24, 2023).
3. https://galaxyproject.org/news/2023-02-23-structured-data-exports-ro-bco/ (accessed Apr. 24, 2023).
4. https://rdmkit.elixir-europe.org/galaxy_assembly (accessed Apr. 24, 2023).
5. S. Soyland-Reyes et al. "Packaging research artefacts with RO-Crate" Data Science, vol. 5, no. 2, pp. 97-138, 2022, http://dx.doi.org/10.3233/DS-210053
6. https://galaxyproject.org/projects/esg/ (accessed Apr. 24, 2023).