

Towards automating FAIR dataset record deployment as a continuous integration process: use of GitHub Actions (GHA) on RDA maDMP metadata*

Miguel-Angel Sicilia , Alberto Ballesteros-Rodríguez , Elena García-Barriocanal , Marçal Mora-Cantalops , and Alberto Nogales 

University of Alcalá, Computer Science Department & Instituto Ramón y Cajal de Investigación Sanitaria (IRyCIS)
{msicilia, alberto.ballesterosr, elena.garciab, marcal.mora, alberto.nogales}@uah.es

February 7, 2026

1 Background

The FAIR (Findability, Accessibility, Interoperability and Reusability) principles aim to enhance the value and impact of research data by ensuring that data can be easily discovered, accessed, integrated with other datasets, and reused for new research purposes. These principles emphasize that both humans and machines should be able to find and use research outputs effectively. The FAIR principles have gained widespread adoption since their publication in 2016 [8]. However, implementation considerations are important as a supplement to these principles for effective data stewardship. Jacobsen et al. [3] proposed a collection of implementation considerations that demonstrate the need to establish clear, organization-wide policies setting standards for elements such as clear licensing, what constitutes “rich” metadata, and how to guarantee provenance and versioning.

While the FAIR principles emphasize automation and machine actionability in accessing research outputs, there is a complementary view in which automation enables FAIR records to be produced and maintained with a consistent level of quality. Underwood [7] reframes the concept of metadata into a “dynamic” context embedded within CI/CD (Continuous Integration/Continuous Deployment) practice. Similarly, Suetake et al. [6] demonstrate how automated metadata validation and testing in workflow sharing systems can significantly improve the reusability of published research outputs. CI/CD is an established practice in the software industry that automates the building, testing, and deployment of software changes, ensuring consistent quality and rapid delivery. When applied to FAIR metadata publishing, CI/CD principles enable automated validation, quality checks, and deployment of metadata records to multiple repositories, ensuring that metadata remains accurate, up-to-date, and compliant with organizational policies throughout the research lifecycle.

2 Context

GitHub Actions (GHA) has gained widespread adoption as a CI tool due to its integration with GitHub [5]. We have used GHA to prototype an approach to continuous metadata publishing. For organizations

*Extended abstract submitted to the 17th International Conference on Current Research Information Systems (CRIS2026), University of the Azores, Portugal, May 20–22, 2026.

that use GitHub as the official repository for automated versioning of metadata, GHA provides a natural extension to integrate the concept of dynamic, quality-controlled metadata publishing and updating.

In our research institute, metadata for datasets originates from multiple sources across different research projects, including REDCap¹ and other project-specific data collection systems. We integrate the dataset definitions specified in each project’s Data Management Plan (DMP) with the operational metadata generated during the research process. This integrated metadata is then uploaded to OpenMetadata², which serves as our internal research repository and provides a centralized catalog of dataset information. GitHub repositories complement this internal infrastructure by serving as the versioned source of truth for validated, publication-ready metadata records that are subsequently deployed to external FAIR repositories.

3 Validating Datasets

The Data Management Plan (DMP) is the document intended to govern and track data produced or updated during a research project. We follow the principles for machine-actionable DMPs [4]; consequently, an actionable DMP is produced adhering to the RDA maDMP schema³. This schema provides the common data model, upon which internal quality policies are layered using the `madmpy` library [1]. The quality validations include: (a) coherence of dataset artifacts with the specifications in the latest version of the actionable DMP and (b) organization-wide policies on metadata completeness and use of terminologies. For example, within our institute, a common quality rule requires the provision of dataset coding using SNOMED CT or other relevant biomedical terminologies, ensuring semantic interoperability and alignment with established domain standards.

As Candela et al. [2] discuss, there is a variety of FAIR-related metrics and assessment tools available. These metrics can be enforced as part of the automated GHA process, such that if new or updated metadata records do not match the organization’s standards for quality, the publishing process fails. This automated enforcement ensures consistent adherence to FAIR principles across all published datasets.

The final actionable metadata is pushed to GitHub using the organization account and tagged with topics to reflect the publication status of the outcomes (these tags can be used to mark datasets as archived or no longer available). Consequently, repositories become the source of truth from which downstream FAIR metadata records are produced.

4 Version update automation

The GitHub `release` event (of type `published`) triggers the rest of the process, whereby the validated metadata (data upload is not yet implemented, as most of the datasets produced are not publicly disclosed due to patient privacy concerns) is transformed and pushed to one or several repositories (in the prototype, Zenodo and Dataverse are the available targets). In the case of Zenodo, an existing GHA integration is already available⁴; for Dataverse⁵, it was implemented by scripting the metadata upload using the official APIs.

The use of GHA to trigger and control metadata publishing across multiple repositories provides several key benefits: (a) consistency, as the same validated metadata is deployed to all target repositories

¹REDCap (Research Electronic Data Capture) is a secure web application for building and managing online surveys and databases for research studies. <https://project-redcap.org/>

²<https://open-metadata.org/>

³The RDA maDMP (Research Data Alliance machine-actionable Data Management Plan) schema is a standardized format for creating DMPs that can be automatically processed and exchanged between systems. <https://github.com/RDA-DMP-Common/RDA-DMP-Common-Standard>

⁴<https://github.com/marketplace/actions/zenodraft>

⁵<https://guides.dataverse.org/en/latest/api/index.html>

without manual intervention or copy-paste errors; (b) auditability, since all deployments are logged and traceable through GitHub’s built-in workflow execution history; (c) efficiency, as a single release action can propagate metadata updates to multiple endpoints simultaneously; and (d) flexibility, allowing the organization to add or remove target repositories by modifying the workflow configuration without changing the underlying metadata validation processes.

The internal processes within the organization determine the publication of a new release whenever the dataset represents a different version, according to semantic versioning conventions. Since the actions can be configured at the organization level using the special `.github` repository, it is possible to enforce organizational procedures as they evolve.

5 Conclusions and limitations

GitHub Actions have been demonstrated as a viable approach to maintaining FAIR metadata consistently and up to date across different repositories, enabling a policy-guided, organization-wide account of metadata quality and richness. The approach has a relatively low technical entry barrier for organizations with varying IT capacities and implements the “continuous metadata” approach at the conclusion of research projects to guarantee FAIR publication of outcomes.

The current prototype works only with publicly accessible metadata and does not provide automated notifications when publishing errors occur. Furthermore, this has only been tested with examples from past projects, and the integration with the internal system, OpenMetadata, is not yet automated. An important limitation is that the automation of the evolution of the internal FAIR policies themselves is not yet implemented, which is required if consistent metadata quality over time is a requirement for the organization.

References

- [1] Ballesteros-Rodriguez, A., Sicilia, M. A., & Garcia-Barriocanal, E. (2025). madmpy: A Python library for creating and validating Data Management Plans. *SoftwareX*, 31, 102215.
- [2] Candela, L., Mangione, D., & Pavone, G. (2024). The FAIR assessment conundrum: Reflections on tools and metrics. *Data Science Journal*, 23(1).
- [3] Jacobsen, A., de Miranda Azevedo, R., Juty, N., et al. (2020). FAIR principles: interpretations and implementation considerations. *Data intelligence*, 2(1-2), 10-29.
- [4] Miksa, T., Simms, S., Mietchen, D., & Jones, S. (2019). Ten principles for machine-actionable data management plans. *PLoS computational biology*, 15(3), e1006750.
- [5] Rostami Mazrae, P., Mens, T., Golzadeh, M., & Decan, A. (2023). On the usage, co-usage and migration of CI/CD tools: A qualitative analysis. *Empirical Software Engineering*, 28(2), 52.
- [6] Suetake, H., Fukusato, T., Igarashi, T., & Ohta, T. (2023). Workflow sharing with automated metadata validation and test execution to improve the reusability of published workflows. *Gigascience*, 12, giad006.
- [7] Underwood, M. (2023). Continuous metadata in continuous integration, stream processing and enterprise DataOps. *Data Intelligence*, 5(1), 275-288.
- [8] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1), 1-9.