How to (Easily) Extend the FAIRness of Existing Repositories Postprint

Authors: Hahnel, Mark, Valen, Dan

Date: 2022-11-16T00:00:00+00:00

Abstract

Data repository infrastructures for academics have appeared in waves since the dawn of Web technology. These waves are driven by changes in societal needs, archiving needs and the development of cloud computing resources. As such, the data repository landscape has many flavors when it comes to sustainability models, target audiences and feature sets. One thing that links all data repositories is a desire to make the content they host reusable, building on the core principles of cataloging content for economical and research speed efficiency. The FAIR principles are a common goal for all repository infrastructures to aim for. No matter what discipline or infrastructure, the goal of reusable content, for both humans and machines, is a common one. This is the first time that repositories can work toward a common goal that ultimately lends itself to interoperability. The idea that research can move further and faster as we un-silo these fantastic resources is an achievable one. This paper investigates the steps that existing repositories need to take in order to remain useful and relevant in a FAIR research world.

Full Text

Preamble

Keywords: FAIR data; Metadata; Interoperability; Repositories; Data curation

Citation: M. Hahnel & D. Valen. How to (easily) extend the FAIRness of existing repositories. Data Intelligence 2(2020), 192–198. doi: 10.1162/dint_a_{00041}

ABSTRACT

Data repository infrastructures for academics have appeared in waves since the dawn of Web technology, driven by changes in societal needs, archiving requirements, and the development of cloud computing resources. Consequently, the data repository landscape exhibits considerable diversity in sustainability models, target audiences, and feature sets. One universal objective links all data repositories: the desire to make hosted content reusable, building on core cataloging principles for economic and research efficiency. The FAIR principles represent a common goal for all repository infrastructures to pursue, regardless of discipline or infrastructure, as reusable content for both humans and machines is a shared aspiration. This marks the first time repositories can work toward a common goal that inherently supports interoperability, making it achievable to advance research further and faster by dismantling silos around these valuable resources. This paper investigates the steps existing repositories must take to remain useful and relevant in a FAIR research world.

Corresponding authors: Mark Hahnel (E-mail: Mark@figshare.com, ORCID: 0000-0003-4741-0309); Dan Valen (E-mail: dan@figshare.com, ORCID: 0000-0002-9479-6438).

---

1. INTRODUCTION

Public access policies, the open access movement, and open data mandates around federally funded research are shaping infrastructure and researcher workflows by creating efficiencies throughout the research process. The surge in public access mandates from foundations, academic publishers, and government funding agencies means we are now discussing "when," not "if," the majority of academic outputs will be openly available and discoverable on the Internet. In the interim, FAIR itself does not prescribe making all data openly available but does ensure funded research is made as open as possible while remaining as closed as necessary (adhering to restrictions such as personally identifiable information). We are witnessing a global push toward open and transparent research to spur innovation, from the European Commission to the Chinese Academy of Sciences to the White House Office of Science and Technology Policy and associated funding agencies, all advancing directives that create a chain effect of open data mandates among global governments and funding bodies.

Foundational, discipline-specific data resources have long provided highly referenced content to the academic community, such as Pangaea for earth and environmental sciences, the Cambridge Crystallographic Data Centre for chemical structures, and GenBank for genomic data. These repositories enable specific disciplines to share relevant research, but their highly curated nature limits the types of content available. Funding agencies seeking greater collaboration opportunities, improved return on research investment, and enhanced transparency in funded research see value in making data available across disciplines by requir-

ing data management plans and protocols for funded researchers to share data. Consequently, general-purpose or "generalist" repositories have emerged to fill a gap by providing avenues for sharing research data either in disciplines lacking special-purpose repository support or for content inappropriate for submission to discipline-specific repositories.

Research published in generalist repositories may be broader but remains equally important to the research process. The FAIR Data Principles, with their fifteen elements striving to make research Findable, Accessible, Interoperable, and Reusable, provide a low barrier to entry for data providers and scholarly communication platforms seeking to maximize research asset reusability. This is crucial as both disciplinary and generalist data repositories must comply with FAIR principles to improve reuse of publicly funded research.

The authors analyzed available technical and policy documentation for ten repositories to assess each platform' s level of "FAIRness," focusing particularly on open API documentation, documented explanations of data curation and metadata markup practices, and clearly defined preservation policies ensuring data availability. Accompanied by a dataset, this practice paper surveyed five subject-specific data repositories and five generalist data repositories not only to assess FAIRness levels but to provide simple recommendations for standardizing the academy' s data repository infrastructure and bolstering interoperability.

---

2. OBJECTIVES

As FAIR data requirements become more explicit, repositories must respond dynamically. The GO FAIR organization highlights that some requirements need human objectification, while others can be checked automatically. For example, "appropriate metadata" is a nuanced requirement that metadata or subject specialist librarians can evaluate, and researchers in similar fields can confirm whether outputs contain sufficient metadata for understanding and reproduction. However, enabling machines to interpret this for every field and subfield is far more complex.

Conversely, machine-readable licenses are simple to implement and verify (as the name suggests), making it unusual for humans to manually query this during curation workflows, such as checking API documentation or landing page HTML. The FAIR principles are intentionally broad and simple to facilitate adoption. Still, surveying generalist and discipline-specific repositories reveals that targeted recommendations leveraging quick technology fixes can ensure interoperability for humans and machines across repository types. To better understand these repository infrastructures' strengths and weaknesses, the authors examined five disciplinary and five generalist data repositories to provide recommendations for seamless interoperability.

---

3. ASSESSMENT

The FAIR principles were created to maximize benefits from global research investments and ensure datasets and other research objects (such as workflows) emerging from traditional science—those not fitting special-purpose, curated repository requirements—are treated as "first-class" research objects no less important to the research enterprise. Reviewing five discipline-specific and five generalist repositories revealed that the newer, decentralized approach to publishing content via generalist repositories and their employed technologies are in some ways more FAIR-friendly than discipline-specific repositories. This may stem from the legacy, specialized nature of disciplinary repositories, but in seeking ways to increase interoperability and ensure published data meets FAIR requirements for both humans and machines, several recommendations emerged based on technological and policy similarities.

These recommendations seek to sharpen implementation choices following FAIR principles, ensuring legacy repositories with community-curated, trusted content can work alongside generalist repositories and data analysis tools such as Kaggle, Gigantum, or R. The recommendations rely on existing technology and policy standards to advance FAIR principles and ensure data are as findable, accessible, interoperable, and reusable as possible.

4.1 Persistent Identifiers

Exposing persistent identifiers (PIDs) as uniform resource identifiers (URIs) is the first recommendation. All generalist repositories explored used data object identifiers (DOIs) for public items from either DataCite or Crossref, Open Researcher and Contributor IDs (ORCIDs) for identifying and disambiguating authors, and Research Organization Registry (ROR) IDs or global research identifier database (GRID) to disambiguate research institutions. Using a central DOI provider ensures public content resolves without link rot or URL changes, indexes all content in a central metadata store (DataCite or Crossref), and ultimately aids discoverability and citation ease. With ORCIDs and ROR/GRIDs, data providers can ensure uniform, searchable provenance and disambiguated PIDs for researchers and institutions, respectively.

These PIDs ensure minimum metadata assignment to published data and digital resources while enhancing discoverability, easing citation, and enabling richer shared metadata across the scholarly community. While any PID suffices, the authors recommend DOIs for immediate integration with current scholarly indexing systems.

4.2 Application Programming Interface (API)

A major differentiator between audited disciplinary repositories and generalist data repositories was API availability. A key tenet of FAIR principles is ensuring published data is understandable for both humans and machines (i.e., programmatically queryable), best achieved by publishing data alongside well-documented APIs.

The OpenAPI Specification provides a standard for machine-readable interface files documenting RESTful API services. Its power lies in being language-agnostic, allowing web clients to understand and consume services without server implementation knowledge or access to server code, and enabling security vulnerability audits of OpenAPI interface files.

Nearly all generalist repositories explored offered well-documented REST APIs ensuring programmatic file and metadata access via URIs, whereas discipline-specific repositories utilized a mix of file transfer protocol (FTP), SOAP API, OAI-PMH, and really simple syndication (RSS) feeds for metadata access. While these options could potentially meet FAIR principles, providing API access adhering to the OpenAPI specification further eases the path to accessibility, interoperability, and data reuse.

4.3 Data Curation and Moderation Workflows

At the technological level, repository providers should enable trained data curators to work in detailed yet efficient manners by providing technical workflows for content review. Data curation is time-consuming and has proven difficult even before introducing technology for large-scale pre-publication collaboration. Communities such as the Data Curation Network have emerged to address training, staffing, and support, yet repository infrastructure has not kept pace in offering technical solutions.

Providing a moderation space where curation experts can check, verify, and approve content, README files for context, and metadata completion is suggested to aid system interoperability. Generalist repositories perform this ad hoc, highlighting a weakness that potentially hinders interoperability and content trust.

4.4 Accessibility

Accessibility in the FAIR context considers technological protocols like HTTP(S), SMTP, and FTP. Indeed, these represent a low bar for ensuring free and open content. These accessibility guidelines focus primarily on W3C standards to ensure open, implementable protocols but do not include Web

Content Accessibility Guidelines (WCAG), making it difficult to achieve a pure level of openness.

4.5 Licenses for Reuse

Clearly defined licenses for both metadata and files are crucial concerning legal reuse rights. These must be human- and machine-readable and assigned to both research files and metadata.

4.6 Sustainability

Preservation workflows ensure PIDs always point to research outputs along with associated metadata. However, as data file sizes grow, repositories must find ways to support themselves long-term. Disaster recovery for repositories should be documented, and integrations with preservation workflows such as Preservica and Archivematica should be shared.

---

5. DISCUSSION

Repository review highlighted high levels of existing commonality regarding core FAIR principles. Technological improvement areas include APIs and accessibility for all, including those with disabilities. System differences mainly concern sustainability models and human curation costs. Most repositories rely on grant-based funding for some costs and do not charge for usage, meaning increased service usage does not automatically yield more scaling budget. This is particularly relevant for mass curation needs.

Most Interoperability and Reusability parameters defined by FAIR uniquely require both human and technological capacity. We recommend repositories add data curation and metadata experts to their teams to enable research building. Both Dryad and ICPSR charge for curation, a model seemingly accepted by the academic community and thus a potential route to improved reusability.

Google Dataset Search achieved low-level interoperability by defining engagement rules for repositories, either through schema.org Dataset markup or equivalent structures in W3C' s Data Catalog Vocabulary (DCAT) format. Google Dataset Search forced a coordinated approach to metadata standards across repositories, and given Google' s near-monopoly on web discoverability, this uniquely drives repositories to collaborate. For the commons' benefit, repositories must also coordinate on preferred metadata schemas and interoperability strategies.

Supporting common querying of subject-specific metadata across diverse subjects is an abstract problem. An explosion in standards means interoperability guidelines must be adhered to for building upon research across multiple repositories. This is acknowledged by only 21% of surveyed repositories reporting their customized vocabularies as fully FAIR. Initiatives such as Metadata 2020 attempt to simplify and standardize metadata assignment to improve research metadata quality. The most underserved FAIR principle aspect in the repository community is metadata quality, volume, and consistency.

Our recommendation is for stakeholders to work toward guidelines through the Research Data Repository Interoperability Working Group—a framework for all repositories to follow if they wish to be FAIR. Similar workflows have been crafted for parallel objectives, such as the Data Policy Standardization and Implementation Interest Group and their recent recommendations.

AUTHOR CONTRIBUTIONS

Both authors M. Hahnel (Mark@figshare.com) and D. Valen (dan@figshare.com) contributed equally to the design and writing of the article. D. Valen created the referenced dataset.

REFERENCES

[1] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, ···& B. Mons. The FAIR guiding principles for scientific data management and stewardship. Scientific Data 3(2016), Article No. 160018. doi: 10.1038/sdata.2016.18.

[2] Guidelines on FAIR data management in Horizon 2020 (2016). Available at: http://ec.europa.eu/research/participants/data/ref/h2020/grants_{manual}/hi/oa_{pilot}/h2020-hi-oa-data-mgt_{en}.pdf.

[3] D. Valen & K. Blanchat. Overview of OSTP responses chart—raw.indd files. (2016). doi: 10.6084/m9.figshare.1522124.v2.

[4] H. Wilcox, D. Baptista & H. Hope. Wellcome' s open access policy review—consultation analysis. (2018). doi: 10.6084/m9.figshare.6887345.v2.

[5] D. Valen & M. Hahnel. Low level repository FAIR overview. (2019). doi:10.6084/m9.figshare.8312408.v1.

[6] M. Downey. Assessing author identifiers: Preparing for a linked data approach to name authority control in an institutional repository context. Journal of Library Metadata 19(1-2)(2019), 117-136. doi: 10.1080/E.C. Friedberg. Good news on the horizon: The open researcher and contributor ID (ORCID). DNA Repair 9(2)(2010), 102. doi: 10.1016/j.dnarep.2009.12.005.

[8] N. Juty, S.M. Wimalaratne, S. Soiland-Reyes, J. Kunze, C.A. Goble & T. Clark. Unique, persistent, resolvable: Identifiers as the foundation of FAIR. Data Intelligence 2(2020), 30–39. doi: 10.1162/dint_a_{00025}.

[9] S. Schwichtenberg, C. Gerth & G. Engels. From open API to semantic specifications and code adapters. In: 2017 IEEE International Conference on Web Services (ICWS), 2017, pp. 484–491. doi: 10.1109/icws.2017.56.

[10] L.R. Johnston, J. Carlson, C. Hudson-Vitale, H. Imker, W. Kozlowski, R. Olendorf, ···& E. Hull. Data curation network: A cross-institutional staffing model for curating research data. International Journal of Digital Curation 13(1)(2018), 125–140. doi: 10.2218/ijdc.v13i1.616.

[11] M. Vans & P. Franks. A blueprint for preserving virtual world cultural heritage using Preservica & custom metadata schema. In: Archiving 2019: Archiving, Preservation, and Access. Available at: https://www.imaging.org/site/IST/Conferences/Archiving/Archiving_{2019}/IST/Conferences/Archiving/A 33f7-4874-aaef-3ccbf6b27461.

[12] E.P. McLellan. Selecting formats for digital preservation: Lessons learned from the Archivematica Project. Information Standards Quarterly 22(2)(2010), 30. Available at: https://groups.niso.org/apps/group_{public}/download.php/4237/IP_{McLellan}*{Selecting*

[13] V. Stathias, A. Koleti, D. Vidović, D.J. Cooper, K.M. Jagodnik, R. Terryn, ···& S.C. Schürer. Sustainable data and metadata management at the BD2K-LINCS Data Coordination and Integration Center. Scientific Data 5(2018), Article No. 180117. doi: 10.1038/sdata.2018.117.

[14] D. Ivanovi, B. Schmidt, R. Grim & A. Dunning. FAIRness of repositories their data: A report from LIBER' s research data management working group. (2019). doi: 10.5281/zenodo.3251593.

[15] I. Hrynaszkiewicz, N. Simons, A. Hussain & S. Goudie. Developing a research data policy framework for all journals and publishers. (2019). doi: 10.6084/m9.figshare.8223365.v1.

Note: Figure translations are in progress. See original paper for figures.