OPINION ARTICLE

Related to other papers in this special issue	3 (p30); 4 (p40); 23 (p230); 5 (p47); 20 (p199); 29 (p285); 11 (p108); 7 (p66)
Addressing FAIR principles	F, A, I, R

How to (Easily) Extend the FAIRness of Existing Repositories

Mark Hahnel[1†] & Dan Valen[2]

[1]Figshare, Crinan Street, London, N1 9XW, United Kingdom

[2]Figshare, Cambridge, MA 02139, USA

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. 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.

† 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

The public access policies, open access movement and open data mandates [1] around federally funded research are helping to shape infrastructure and researcher workflows by creating efficiencies throughout the research process. What the jump in public access mandates from foundations, academic publishers and government funding agencies means is that we are now talking about "when", not "if", and the majority of academic outputs will live openly and be discoverable on the Internet. In the interim, FAIR [2] itself does not prescribe to making all data openly available but does ensure funded research be made as open as possible, while remaining as closed as necessary (adhering to restrictions such as personally identifiable information) [3]. We are seeing a global push toward open and transparent research to spur innovation, from the European Commission to the Chinese Academy of Sciences to White House Office of Science and Technology Policy and funding agencies therein, all pushing ahead with directives that are also causing a chain effect of open data directives among global governments and funding bodies.

There are foundational, discipline-specific data resources that have provided highly referenced content to the academic community such as Pangaea[①] for the earth and environmental sciences, the Cambridge Crystallographic Data Centre for chemical structures[②], and GenBank for genomic data[③]. These repositories provide a way for specific disciplines to share relevant research, but the highly curated nature limits the types of content available. Aforementioned funding agencies looking for more opportunities for collaboration, more return on research investment, and more transparency in funded research see value in making data available across disciplines by requirement data management plans and protocols for funded researchers to share data [4]. As such, there has been a rise in general purpose or "generalist" repositories to fill a gap in providing a way for research data to be shared either in disciplines that are not supported by a special-purpose repository or even those that do but have generated content that is not appropriate for submission to that repository.

The research published in generalist repositories may be broader but is just as important to the research process. The FAIR Data Principles, 15 elements of which strive to make research Findable, Accessible, Interoperable and Reusable, provide a low barrier to entry for data providers and scholarly communication platforms alike to consider in making their research assets maximally reusable. This is important as both disciplinary and generalist data repositories have an obligation to comply with the FAIR principles in order to improve reuse of publicly funded research.

The authors analyzed available technical and policy documentation for 10 different repositories to break down the levels of "FAIRness" of each platform. Of special interest were open API documentation, documented explanation of data curation and metadata mark-up practices, and clearly defined preservation policies to ensure data availability. With the accompanied data set [5] this practice paper surveyed five subject specific data repositories and five generalist data repositories in an attempt not only to assess the levels of FAIR-ness, but to provide a number of simple recommendations to help standardize the academy's data repository infrastructure and bolster interoperability.

① https://www.pangaea.de/.
② https://www.ccdc.cam.ac.uk/.
③ https://www.ncbi.nlm.nih.gov/genbank/.

2. OBJECTIVES

As requirements around FAIR data become more explicit, repositories will need to respond in a dynamic manner. The GO FAIR organization[④] highlights that some of the requirements need some kind of human objectification, and some could be checked automatically. For example, "appropriate metadata" is a very nuanced requirement. A metadata or subject specialist librarian may be able to determine this. Someone working in a similar field may be able to confirm that the research output has "all of the metadata required to understand and reproduce the research." However, for machines to be able to interpret this for every single field and subfield of research is a much more complex task.

Conversely, machine readable licenses are a simple thing to implement, and a simple thing for a machine to check for (as the name suggests). It would be odd for a human to query this in a curation workflow, i.e., to check the API documentation or even landing page HTML.

The FAIR principles are broad and simple by design to allow for ease of adoption. Still, in surveying the landscape of generalist and discipline-specific repositories, it has become apparent that there are simple, more targeted recommendations that can leverage quick technology fixes to ensure interoperability for both humans and machines between generalist data repositories and their special purpose counterparts. In an effort to better understand the strengths and weaknesses of these repository infrastructures, the authors explored five disciplinary data repositories and five generalist data repositories to provide recommendations for seamless interoperability.

3. ASSESSMENT

The FAIR principles were created as a way to extract the maximum benefit from global research investments and ensure that data sets and other research objects (for instance workflows) emerging from traditional science that do not fit the requirements of special-purpose, curated repositories are treated as "first class" research objects that are no less important to the research enterprise[6]. In reviewing five different discipline specific repositories and five generalist repositories, it became clear that the newer, decentralized way of publishing content via generalist repositories and the technologies those repositories employ are in some ways more FAIR-friendly than the discipline-specific repositories. This could be due to the legacy, specialized nature of disciplinary repositories, but in looking to ways to increase interoperability and ensure published data meets FAIR requirements for both humans and machines, a number of recommendations emerged based off similarities in technologies and policies.

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

④ https://www.go-fair.org/fair-principles/.

4. RECOMMENDATIONS

4.1 Persistent Identifiers

Exposing persistent identifiers (PIDs) as uniform resource identifiers (URIs) is the first recommendation. All of the generalist repositories explored used data object identifiers (DOIs) for all public items from either DataCite[5] or Crossref[6], Open Researcher and Contributor IDs (ORCIDs)[7] for identifying and disambiguating authors, and Research Organization Registry (ROR) IDs[8] or global research identifier database (GRID)[9] to disambiguate research institution. By using a central DOI provider, it ensures that public content is resolved in a way that prevents link rot or the changing of uniform resource locators (URLs), that all content is indexed in a central metadata store (the aforementioned DataCite or Crossref), and ultimately aid in the discoverability and ease of citation or reference of published content. With ORCIDs and ROR/GRIDs, data providers can ensure a level of uniform, searchable provenance and disambiguated PIDs for researchers and institutions, respectively [6,7,8].

These PIDs ensure a minimum amount of metadata assignment to published data and other digital resources while also enhancing discoverability, easing citation, and ensuring richer, shared metadata by the scholarly community. Whilst any PID will suffice, in order to get more immediate integration in current scholarly indexing systems, the authors recommend the use of DOIs.

4.2 Application Programming Interface (API)

A major differentiator between the audited disciplinary repositories and the generalist data repositories was the existence of an API. One of the key tenets of the FAIR principles is ensuring published data is understandable for both humans and machines (in this case, query-able programmatically), and the best way to ensure this publishing data alongside a set of well-documented APIs.

The OpenAPI Specification is a standard for machine-readable interface files for documenting RESTful API services. What is so powerful about the OpenAPI specification is that it is language-agnostic, it allows Web clients to understand and "consume" services without knowledge of server implementation or access to the server code, and ensure OpenAPI interface files can be audited for security vulnerabilities[10] [9].

Nearly all of the generalist repositories explored had a well-documented REST API, ensuring the programmatic access to files and metadata via URIs, whereas the discipline-specific repositories utilized a mix of file transfer protocol (FTP), SOAP API[11], OAI-PMH[12], and really simple syndication (RSS) feeds as a

[5] https://datacite.org/.
[6] https://www.crossref.org/.
[7] https://orcid.org/.
[8] https://ror.org/.
[9] https://grid.ac/.
[10] https://en.wikipedia.org/wiki/OpenAPI_Specification.
[11] Simple Object Access Protocol (SOAP) is protocol that defines a uniform way of passing XML-encoded data.
[12] OAI-PMH is a protocol for exposing the structured metadata of a data repository.

way to access metadata. While these options above could potentially meet the FAIR principles, going a step above by providing API access that adheres to the OpenAPI specification eases the route to accessibility, interoperability and data reuse.

4.3 Data Curation and Moderation Workflows

At the technological level, repository providers should ensure that trained data curators can work in a detailed yet efficient manner. Repository providers should provide technical workflows within their systems to allow data curators to review.

Data curation can be a time-consuming process on its own and has proven difficult even before you introduce the technology needed for large-scale collaboration prior to publication. There are a number of communities that have sprung up to address training, staffing and support of data curation such as the Data Curation Network® [10], yet repository infrastructure has not kept pace in offering technical solutions.

Providing a moderation space within a platform where curation experts can check, verify, and approve content, README files for context, and metadata completion is suggested to aid interoperability between systems. This is performed ad hoc among generalist repositories which highlight a weak spot, in turn potentially hindering interoperability and trust of content.

4.4 Accessibility

Accessibility as it relates to FAIR takes technological protocols like HTTP(S), SMTP and FTP into account. Indeed, these are a low bar to ensure content is free and open. These accessibility guidelines focus mainly on W3C standards to ensure the protocol is open and implementable. However they do not include Web Content Accessibility Guidelines (WCAG). Without these guidelines, it is difficult to achieve a pure level of open.

4.5 Licenses for Reuse

Clearly defined licenses for both metadata and files are of utmost importance when it concerns legal rights of reuse of content. These must be human and machine readable and assigned to both the actual research files as well as the metadata.

4.6 Sustainability

Preservation workflows ensure that PIDs should always point to the research output in question, along with the associated metadata. However, as data file sizes grow, repositories must find a way to support themselves in the long term.

® https://datacurationnetwork.org/.

Disaster recovery for the repository should be documented and shared integrations with preservation workflows such as Preservica [11] and Archivematica [12].

5. DISCUSSION

A review of repositories highlighted high levels of existing commonalities with regard to core FAIR principles. On a technological level, areas for improvement include APIs and Accessibility for all, including those with disabilities. Areas where these systems do differ mainly cover sustainability models and human curation costs. The majority of the repositories rely on grant-based funding for some of their costs and do not charge for usage. Thus, increased usage of the service does not automatically mean more budget available for scaling. This is particularly relevant with regard to the need for higher curation en masse.

The majority of the Interoperability and Reusability parameters defined by FAIR are unique in that they need both human and technological capacity. We recommend that repositories work to add data curation and metadata experts to their teams in order to allow research to be built upon. Both Dryad and ICPSR charge for curation, a model that seems to be accepted by the academic community and therefore a potential route to improved reusability.

Interoperability has been achieved at a low level by Google Dataset Search[®] by defining rules of engagement for repositories, either schema.org Dataset mark-up, or equivalent structures represented in W3C's Data Catalog Vocabulary (DCAT) format. What Google Dataset Search has forced is a co-ordinated approach to metadata standards across repositories. As Google has a near monopoly of Web discoverability, this is a unique driver for repositories to work together. For the good of the commons, repositories must also coordinate on preferential metadata schemas and interoperability strategies.

Trying to support common querying of subject specific metadata across diverse subjects is an abstract problem. An explosion in the number of standards [13] means that interoperability guidelines need to be adhered to in order to build on top of research across multiple repositories. This is acknowledged with just 21% of surveyed repositories commenting that their customized vocabularies are fully FAIR [14]. There are several other initiatives, namely Metadata 2020[®], who are attempting to simplify and standardize metadata assignment to improve the quality of metadata for research. The most underserved aspect of the FAIR data principles in the repository community is quality, volume and consistency of metadata.

Our recommendation is for interested stakeholders to work toward such guidelines through the Research Data Repository Interoperability Working Group – a framework for all repositories to adhere to, should 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 [15].

[®] https://toolbox.google.com/datasetsearch.
[®] http://www.metadata2020.org/.

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 data set.

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: 10.1080/19386389.2019.1590936.

[7] 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/Archiving2019/Archiving_2019_Home.aspx?hkey=54bdc838-33f7-4874-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_Formats_isqv22no2.pdf.

[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.