Open Science Monitoring

Impact Case Study – Data FAIRport
Author information and acknowledgements

This study was delivered by RAND Europe, with the support of Deloitte and Digital Science & Research Solutions, including Altmetric.com and figshare. The project was commissioned by the European Commission Directorate-General for Research and Innovation and received funding under specific contract 2015/RTD/A6/SC/PP-02741-2015.¹

The team at RAND Europe included: Elta Smith (project leader), Salil Gunashekar (project manager), Sarah Parks, Catherine Lichten, Louise Lepeit, Molly Morgan Jones, Catriona Manville and Calum MacLure. The team at Deloitte included: Valentina Cilli, Maruca De Ramón, Vilma Zotou, Benoit Vandresse and Sebastiaan van der Peijl.

Data sources

We would like to thank the following organisations for generously contributing their data for use on the monitor: 101innovations, Altmetric, arXiv, bioRxiv, Clarivate Analytics, F1000Research, figshare, GenBank, Helmholtz-Centre for Environmental Research – UFZ and German Centre for Integrative Biodiversity Research (iDiv), Nature Publishing Group and Palgrave Macmillan, OpenAIRE, PeerJ preprints, Publons, re3data.org, RePEc, SHERPA/RoMEO, SHERPA/JULIET, Taylor & Francis, and Wiley.

Expert panel

We are grateful to our expert panel members who provided helpful advice and guidance: Aletta Bonn (German Centre for Integrative Biodiversity Research (iDiv) and Helmholtz Center for Environmental Research (UFZ), Friedrich-Schiller-Universität Jena, Germany); Thomas Crouzier (School of Biotechnology, Royal Institute of Technology (KTH), Sweden); Ivo Grigorov (National Institute of Aquatic Resources, Technical University of Denmark); Peter Kraker (Postdoctoral Fellow, Know-Center Social Computing Group, Graz University of Technology, Austria); Eleni Malliou (National Documentation Center, Greece); and Paul Wouters (CWTS-Centre for Science and Technology Studies, Leiden University, The Netherlands).

Acknowledgements:

We would like to thank the following individuals for their contributions to the monitor: Susan Guthrie (RAND Europe), Gavin Cochrane (RAND Europe), Alexandra Pollitt (King’s College London), and Marion Bywater who acted as our quality assurance reviewers and provided thoughtful and timely comments on our analyses; Jessica Plumridge (RAND Europe) who designed the graphics presented in the online monitor; Monica Hertzman (RAND Corporation) who provided technical assistance in relation to the design of the website; Steven Wooding (RAND Europe) who provided advice and support throughout the study; Vincent Larivière (Observatoire des Sciences et des Technologies) for his insights and advice; Joanna Chataway (University of Sussex) for her advice and support; Rebecca Ioppolo (RAND Europe) who provided research assistant support; Jack Melling (RAND Europe) who helped in the dissemination activities related to the launch of the monitor; and the many individuals and organisations who participated in an online consultation we conducted as part of the development of the monitor.

For more information about this document or the Open Science Monitor, please contact:

Elta Smith
RAND Europe | Westbrook Centre, Milton Road, Cambridge CB4 1YG, United Kingdom
Telephone: +44 (1223) 353 329 | E-mail: eltas@rand.org

Data FAIRport

Summary

The Data FAIRport initiative (2017) aims to provide a minimal but comprehensive framework for developing and implementing good management and stewardship of research data and metadata in the life sciences. The initiative does not suggest any technology or protocol to achieve this goal, instead it provides a set of guiding principles and a framework to make research data findable, accessible, interoperable and re-usable (FAIR).

Started in January 2014, the Data FAIRport initiative aims to support the life sciences research community in reconciling the increasing volume of research data that is produced with the ability to analyse and link the data, because the latter has not developed at the same speed.

The FAIR guiding principles were developed and are implemented in an open and collaborative manner by the life sciences and data research community. Open and collaborative methods are often used to develop technological solutions such as ‘hackathons’ and ‘bring-your-own-data’ (BYOD) parties.

Background

The life sciences are inherently data-intensive disciplines that collect and produce large volumes of data at several levels – from the molecular level, to the organism level, to populations as a whole – as well as interactions among those. The volume of data produced by research has grown exponentially over the years. Key factors in this growth have been technological developments that have increased computational and storage capacity and the speed of data transport. The volume of data has grown more rapidly than the capacity to analyse and link the data, however, resulting in duplication of effort and reducing the potential for researchers to reuse or connect datasets that are relevant for their research (Data FAIRport 2014a). Researchers have focused increasingly on the need for data stewardship to resolve these issues; funders and donors have also begun to require sound data stewardship plans as a condition of awarding research grants. Data stewardship is at the heart of the Data FAIRport initiative, which emphasises that stewardship is more than just good data management. It includes ‘long-term’ care of valuable digital assets to ensure that they are accessible and available for reuse over time (Data FAIRport 2017).

The Data FAIRport initiative started in January 2014 as a follow-up to the 'Jointly Designing a Data FAIRport' workshop, organised in Leiden, the Netherlands (Data FAIRport 2014b). During the workshop, 25 life sciences researchers and experts discussed the need for a global infrastructure for publishing, discovery, exchange and reuse of life sciences research data. The event was organised at the request of the Netherlands eScience Center (2017)

---

2 For example, the Netherlands Organisation for Scientific Research (known by the acronym NWO) announced in 2013 that it would be changing the terms for assessing grant-supported researchers. The new criteria would include not only the impact of their publications but also the ‘products’ of the research or research data. Therefore, if a researcher had more citations for a dataset, video, or snippet of code than for a traditional article, it could be considered more impactful.

3 Workshop participants included representatives from leading research infrastructures and policy institutes, publishers, semantic web specialists, innovators, computer scientists and experimental scientists.

4 eScience stands for experimental science.
and the Dutch Techcentre for Life Sciences (DTL) (2017), which are among the organisations most active in the definition and implementation of the FAIR principles.

The Data FAIRport initiative aims to provide a minimal but comprehensive framework in which issues in data discoverability, access, annotation and authoring can be addressed. The initiative does not aim to create new technologies, standards or protocols, but, rather, to identify and promote the use of minimal requirements for data to be findable, accessible, interoperable and reusable.

**Methods and the role of open science**

The vision of the Data FAIRport initiative is 'to join and support existing communities that try to realise and enable a situation where valuable scientific data is 'FAIR’ in the sense of being Findable, Accessible, Interoperable and Reusable' (Data FAIRport 2017).

A group of the stakeholders drew up the FAIR guiding principles based on the four core characteristics; these principles are described on the Data FAIRport website (2016). The principles were initially proposed at the workshop in 2014, and they have been discussed and refined through the contributions of different stakeholders over time (Data FAIRport 2016). They are related but independent and separable, and they define characteristics that data resources, tools, vocabularies and infrastructures should have to enable discovery and reuse by third parties. The FAIR guiding principles define minimum requirements and are modular to support different circumstances (e.g. highly sensitive or personal data) based on the principle that the availability of a rich set of metadata provides a high degree of ‘FAIRness’, even in the absence of FAIR publication of the data itself.

The FAIR guiding principles do not require any specific technology, standard or implementation protocol, nor are they themselves a standard or specification. They assist researchers and data publishers in assessing whether their implementation choices make their research data findable, accessible, interoperable and re-usable.

The scope of the initiative involves a set of elements to which the FAIR guiding principles apply, including (Starr et al. 2015):

- Adoption and communication of standards
- Simplification of standard interoperation
- Adoption of cross-cutting standards for origin, versioning, identity and dependency for data and metadata covering identifiers, formats, checklists and vocabularies
- Interoperation of data services
- Reconciliation of evolving standards and the datasets organised or annotated by them
- Minimal models of investigation for grouping results
- Metadata required to link data with analytics
- Data citation (mechanics, adoption, recognition)

---

5 A document that reports the community discussion relating to clarifications and explanations of these principles, and detailed guidelines and examples of FAIR implementation, is being implemented.

6 The Data FAIRport initiative adopts the ‘hourglass’ approach, which defines the absolute minimum that is required to achieve interoperability (similar to the approach that defines the Internet and other robust, heterogeneous but interoperable infrastructures).
While it does not implement specific projects directly, the Data FAIRport initiative relies on other projects, organisations and groups for the implementation of its principles and guidelines in research. As such, it actively promotes the development of an open community of 'enablers' using and publishing findable, accessible, interoperable and re-usable data.

**Outputs and findings**

The groups of stakeholders (at least potentially) interested in the Data FAIRport initiative are numerous and heterogeneous (Starr et al. 2015). They include:

- Researchers wanting to share and reuse data and analyses carried out in other relevant studies
- Professional data publishers willing to offer better services
- Software and tool builders providing data analysis and processing services, such as re-usable workflows
- Funding agencies (public or private) increasingly demanding long-term management and stewardship of data
- The data science community dealing with mining, integration and analysis of an exponentially growing set of data

The group of workshop participants has grown over time, from 25 members initially to 32 members in 2017 from different backgrounds and from all the relevant stakeholder groups. The refinement of the FAIR data guiding principles is managed by an ad hoc group within the FORCE11 community, an open community of scholars, librarians, archivists, publishers and research funders that aims to support the move towards modern scholarly communications through the effective use of information technology.

Several repositories are being developed in accordance with the FAIR guiding principles, using a variety of technological solutions. Relevant examples are provided below.

- **Dataverse** (Dataverse 2017) is open source data repository software, installed in many research institutions around the world to support public community repositories or institutional research data repositories. Dataverse generates a formal citation for each deposit and provides a landing page that grants access to metadata, data files, dataset terms, waivers or licences, and versioning information, all of which are indexed and searchable. Metadata is always public, even if the data are restricted or removed for privacy issues.

  Dataverse provides public, machine-accessible interfaces to search the data, access the metadata and download the data files, using a token to grant access when data files are restricted. The Dataverse software supports the four FAIR data principles because metadata (and data, if not restricted) are findable, accessible, interoperable and re-usable. Harvard Dataverse, with more than 60,000 datasets, is the largest of the current Dataverse repositories, and it is open to all researchers from all research fields (Harvard Dataverse 2017).

- **FAIRDOM** (2017) is a project aiming to establish a data and model management service facility for the field of systems biology based on the FAIR data principles. Individual research assets (or aggregates of data and models) are identified with unique and persistent HTTP URLs, which can be registered with digital object

---

7 The FORCE11 community currently has almost 2,000 members and 31 work groups developing principles and tools for better data management and stewardship in eScience (Force11 2016).
identifiers (DOIs\(^8\)) for publication. Research assets can be accessed over the Internet in a variety of formats (e.g. RDF, XML) and are annotated with rich metadata, using community standards, formats and ontologies. The metadata are stored as an RDF to enable interoperability and downloading of assets for reuse.

- **Open PHACTS** (2015) is a data integration platform for information pertaining to pharmacology based on the FAIR data principles. An interface provides access to the platform (via multiple formats, both human- and machine-readable) and allows multiple URLs to be used to access information about a particular asset through a mapping service linked to pharmacology databases (such as Chemsper or DrugBank). All data sources used are described using standardised dataset descriptions. Finally, a majority of the datasets are described using community-agreed ontologies.

- **UniProt** (2017) is a resource for protein sequence and annotation data that is consistent with the FAIR data principles. All entries are uniquely identified by a stable URL, which provides access to the record in a variety of formats, including a web page, plain text, and RDF. The record contains rich metadata that is both human- and machine-readable. UniProt links with more than 150 research databases (such as PubMed), so that each UniProt record has links to such databases, enabling rich citation.

In addition to, and in support of, communities already implementing the FAIR data principles, Force11’s Data Citation Synthesis Group has published the *Joint Declaration on Data Citation Principles* for direct data citations in scholarly articles, and the Data Citation Implementation Group and the Data Citation Implementation Pilot have started to work on their practical implementation (Martone 2014). The *Joint Declaration* has received more than 350 endorsements by different organisations (from universities and research institutes to networks of libraries and publishers).\(^9\)

Several projects developing data infrastructures, semantics and data mining techniques for research aim to achieve their objectives with the Data FAIRport principles, such as ODEX4All (DTL 2015), Elixir (2017), Force11 (2016), BD2K (NIH 2016), Research Data Alliance (RDA 2017), ENPADASI (JPI HDHL 2017), and BBMRI-NL (2017).

In addition, working groups have been created within specific project teams (termed skunk working groups or skunkworks groups (Wilkinson 2015) to develop software-supporting infrastructures that are end-to-end, are compatible with the FAIR principles, and can be implemented through existing repositories. Such FAIR-related activities are organised in the form of open, collaborative methods such as hackathons\(^10\) and/or BYOD parties.\(^11\) During 2016, both the Elixir and the ODEX4All projects have organised BYOD parties to implement

---

\(^8\) The DOI is a standard used to persistently and uniquely identify objects. The DOI system is particularly used for electronic documents, such as journal articles. The DOI system started in 2000 and is managed by the International DOI Foundation. See: https://www.doi.org/

\(^9\) For the full list of endorsements, see Force11 (2016).

\(^10\) A hackathon is a design-sprint-like event of variable duration (in general lasting from one day to one week) in which computer programmers and software developers collaborate intensively on a software project, often together with subject matter experts (Leckart 2012).

\(^11\) A bring-your-own-data party is a collaborative event (usually lasting from one to two days) in which subject-matter experts work together use with programmers and software developers on their own datasets to make them interoperable. The event includes a combination of preparatory works (such as webinars), hackathons and tutorials. At the end of a BYOD party, participants should be able to design their datasets in accordance with the FAIR principles (DTL 2016b).
the Data FAIRport principles. In addition, the Dutch Techcentre for Life Sciences, one of the main supporters of the Data FAIRport initiative from the outset, has organised several hackathons (DTL 2017c) and BYOD parties (DTL 2017b) to support the implementation of the FAIR guiding principles in 2016 in the life sciences domain.

The work on the Data FAIRport initiative has generated a number of academic publications, including Roos et al. (2016), Starr et al. (2015) and Wilkinson et al. (2016).

Impacts

An increasing number of projects and research communities are working with the FAIR principles, and they often do so using open and collaborative methods, such as hackathons and/or BYOD parties.

Sources


———. 2014b. ‘Data FAIRport Conference: Jointly Designing a Data FAIRport.’ As of 19 January 2017: http://www.datafairport.org/component/content/article/8-news/9-item1


