# Making FAIR Practices Accessible and Attractive

Lyudmila Balakireva(✉) and Fedor Balakirev

Los Alamos National Laboratory, Los Alamos, NM 87545, USA
ludab@lanl.gov

**Abstract.** Facing rapidly growing volumes of research datasets, scientists and research funding agencies are putting forward new principles of data management, such as data Findability, Accessibility, Interoperability, and Reusability (FAIR). To this end, data science experts are developing FAIR data policies, methods, protocols, and repositories, while actual research practices are lagging behind because FAIR compliance remains a burden for many researchers. Here we present a prototype data management infrastructure deployed at the National High Magnetic Field Laboratory (NHMFL/MagLab) aimed at helping scientists efficiently annotate and manage experimental data produced by their MagLab projects and making FAIR practices accessible and attractive. The infrastructure incorporates the Open Science Framework (OSF) data repository platform. We will describe infrastructure elements such as the data formats, the metadata schema, the repository integration, the naming conventions, the templates to organize the data, and the automated data pipeline from measurement stations to the FAIR repository objects.

**Keywords:** FAIR data · Open science · Repository · Dataset publishing

## 1 Introduction

With the advances in machine learning technology and the availability of vast quantities of digital information, data are becoming more usable and valuable than ever. At the same time, only properly annotated data objects can be reused by humans and machines to gain viable scientific output [34]. Alongside the persistence and accessibly of the data object, the information that puts it in a context of a research field and outlines its purpose is equally important. Many of the ideas in the FAIR principles gear towards enabling "machine actionability", or automatic knowledge extraction, where data object identifying information plays a significant role. Can we also automate collecting of the data and metadata into data objects which are FAIR-ready and immediately useful? Can we build

Supported by the Department of Energy, the National Science Foundation Cooperative Agreement No. DMR-1644779, and the State of Florida.


G. Silvello et al. (Eds.): TPDL 2022, LNCS 13541, pp. 417–424, 2022.
https://doi.org/10.1007/978-3-031-16802-4_41

upon established research tools and practices instead of forcing scientists to start from scratch in order to comply with FAIR standards?

While major sponsors of research strongly encourage the implementation of the FAIR data management principles [35], data acquired in a laboratory is often not annotated in an informative way, e.g. handwritten rather than electronic lab notebooks, or poorly described and widely varying data file formats. Even if made public, such datasets are often cannot be utilized without some extrinsic contextual knowledge, such as a separate scholarly communication. According to a prior MagLab study [31] of data management problems in the condensed matter physics community, the most often noted problem is the lack of file naming conventions (52%), followed by the difficulties in interpreting data due to poor or lost documentation (50%); the lack of version control (36%); and the inability to access data due to obsolescence, proprietary formats, expired software licenses, or other issues (35%). It is clear that FAIR compliance will require major changes in the research culture and the implementation and normalization of data management technologies and practices [28,30,35].

The MagLab Pulsed Field Facility at the Los Alamos National Laboratory [27] attracts hundreds of researchers from around the world to conduct fundamental and applied science experiments at the highest possible magnetic field intensity. The diverse environment of the MagLab user program allows for a comprehensive study how the practices and methods of motivating scientists to engage with the FAIR principles might be changed for the better. We studied all the stages of the dataset lifecycle, and we demonstrate how FAIR practices and automatic data object creation could be introduced in such an environment in an approachable and efficient fashion.

In this paper, we will describe the research data management infrastructure devised and implemented at the MagLab, as outlined on Fig. 1. We propose the data object creation process which links the description of the research project with the metadata-rich experimental results. The project description provides scientific context and purpose of the data object, while the proposed hierarchical data format allows for data organization and for metadata to be included at different layers of the composite object. The resulting dataset files are encapsulated into FAIR-ready data object with sufficient information for its reuse. We also show how to increase the degree of automation for the collection, aggregation, and repository submission of the scientific data that is ready for sharing, where the repository project itself can become FAIR data object and actionable knowledge unit.

## 2 Related Work

Because experimental science needs flexibility to change the methodology and the parameters, automating data management workflows at the time of creation can assure that FAIR principles are upheld from the outset [29], keeping data objects stable. Several scientific communities started designing and implementing solutions in this direction such as in climate modelling [7], language data and

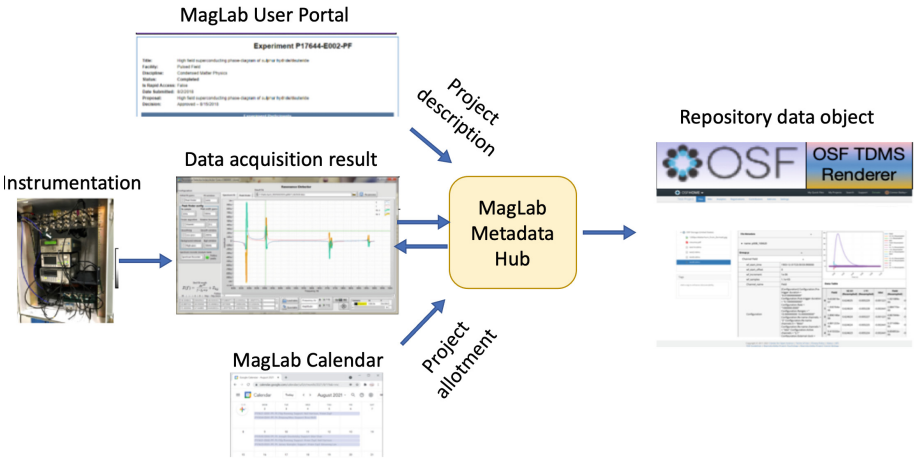


**Fig. 1.** Architecture of the FAIR-ready data management framework across heterogeneous systems. Data recorded during the user experiment is automatically uploaded to the corresponding OSF project in FAIR-ready format.

technologies [5] and material science [21]. Online FAIR data repositories [6] help facilitate the adoption of FAIR standards by providing frameworks for storage, access, search, Application Programming Interface (API), and other features that create organized hubs of scientific data. One needs to distinguish between the notions of FAIR data, open data access, and data management. While data management is about the stewardship of data from the cradle onwards, open data focus on access, and FAIR goes beyond with promotion of truly reusable data [13]. For example, the FAIR repository standards can even be adhered to in the restricted environments [4,9] with locally installed OSF [23] as secure internal data management and collaboration space for LANL research. Switching to electronic and open lab books can also ease FAIR transition, where publicizing the details of the experimental dataset creation can help establish its context and purpose [21,29].

## 3 Results

### 3.1 Experimental Data Set Acquisition, Formats, and Metadata Descriptors

MagLab implements a uniform data acquisition (DAQ) infrastructure to record experimental data at each of the MagLab DAQ stations [12]. The DAQ infrastructure utilizes National Instruments LabVIEW software [1,15] incorporating a wide variety of DAQ instruments. Technical Data Management Streaming (TDMS) format [32] was selected as the primary standard for the data records. TDMS is an industry-standard structured open data format with flexible metadata options to describe the data at each level of the hierarchy. Structured data

formats, including TDMS and Hierarchical Data Format Five (HDF5) [11] are finding broader acceptance in scientific communities. The structured formats address FAIR interoperability because it is supported by many scientific software packages, such as MATLAB [20], Origin [24], NumPy [22], IGOR [14], Excel [8], and it can contain rich metadata, making data usable to others. The metadata can be incorporated at every level of the hierarchy to help explain and annotate the structure, contents, and format of the data. For example, such acquisition metadata as the DAQ parameters and information describing how the data should be read can be embedded at the same level as the raw data.

### 3.2 Dataset-Identifying Descriptors

Any external or local researcher (user) can submit a research proposal to request access to the MagLab facilities at the User Portal [17], thus creating the first descriptive metadata about the project, including project and grant identifiers, experiment purpose and description, primary investigator (PI), DAQ station, and the dates of the experiment. This type of the "administrative" metadata addresses contextual questions such as "who", "when", "where" and "why". The administrative metadata is important to preserve together with the acquisition metadata, which describes "how" the data was generated.

The metadata that identifies the proposal exists as a separate administrative record kept by the MagLab headquarters [17], which was not accessible to the DAQ software. To address this gap, we developed a Metadata Hub [19] that serves as a linking element to collect proposal metadata from existing administrative sources and provide Representational State Transfer (REST) back-end API with easily parsable JavaScript Object Notation (JSON) format [3]. Each DAQ station can now request information about the scheduled experiment from Metadata Hub. We also developed a web-based front-end to aid local administrators and support staff scientists in tracking MagLab proposals and their corresponding metadata.

We enhanced the DAQ software with a Project Metadata dashboard describing the experiment (see Fig. 2). The dashboard displays the key proposal identifiers, such as the PI and the proposal title. If the displayed proposal does not match the actual one (e.g. researcher's experiment was relocated to a different DAQ station), the user can select the correct proposal from the list.

### 3.3 FAIR Data Repository Integration

MagLab scientists survey pointed out that while most scientists do not use repositories, there is an occasional usage of inconsistent sharing solutions. A common repository platform for MagLab users and uniform data curation standards will facilitate data accessibility and reuse. Standardization also allows for automation of dataset archiving and retrieval. An online repository provides additional benefit for a dispersed team, where an efficient data pipeline and shared repository space shorten the time from data production to analysis.

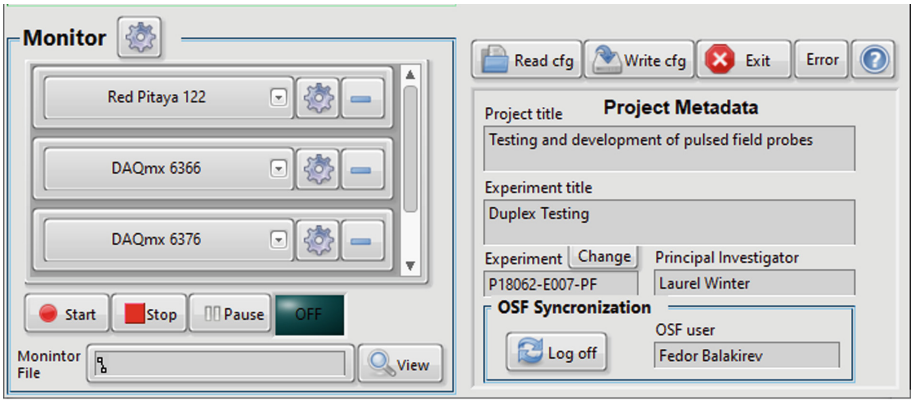


**Fig. 2.** Fragment of MagLab DAQ Software interface with new Project Metadata dashboard with the project identification information and OSF Synchronization section.

Repositories must earn the trust of the communities they intend to serve [16], while providing persistent and dependable services. OSF [10] is a data repository that seeks to facilitate open collaboration in scientific research by combining many attractive features including structured projects, ample data storage, and integration with third-party services. To streamline data sharing and access, as well as eventual publication of the datasets, the MagLab has selected the Open Science Framework as the repository standard.

The users willing to publicize their experiment as a persistent data object can accomplish the task with a push of the OSF login button. Once the user authenticates to OSF via OAuth2 mechanism [19,25,33], the Metadata Hub automatically creates a project at the OSF repository and pre-populates it with the proposal metadata that was harvested from the User Portal. The Metadata Hub also assigns several common tags to the OSF projects, such as "maglab" and project identifier. The use of common tags to find the MagLab projects also facilitates discoverability and has the potential to showcase all achievements of the MagLab to the public. We created a web TDMS renderer tool as OSF plug-in [2]. By bringing this user-friendly tool to a popular FAIR data repository, we hope that we have helped to foster a FAIR-ready environment for scientists.

The data recorded during the user experiment is automatically uploaded to the corresponding OSF project via HTTP protocol. For each uploaded TDMS binary file, we create a companion ASCII markup file which lists the metadata extracted from the TDMS file to enhance the data findability. We also provide an "electronic lab book" functionality as part of the dataset synchronization, where a wiki-based OSF electronic lab book is pre-populated with the file metadata to help users track, contextualize, and share the experimental details using OSF file-sharing capabilities.

### 3.4 Data Management Architecture

The overall data management architecture is summarized in Fig. 1. The Metadata Hub periodically updates its database with administrative metadata from the MagLab calendar [18] and User Portal maintained by the headquarters. The user sets the experimental parameters at the DAQ station computer. If the user grants the permission to synchronize the data collection with the OSF, the Metadata Hub creates the project space and pre-populates it with the proposal metadata.

As the user experiment captures the data, the Metadata Hub uploads new TDMS data files to the OSF project, and updates the electronic lab book. The user can add their own comments and experimental details to the lab book at the OSF portal. The user scientist can also share the OSF project with their collaborators. The initial OSF project is private and is only shared with project collaborators. The user can publicize the project and create its persistent digital object identifier (DOI).

### 3.5 Discussion and Future Work

The adoption of FAIR standards remains a challenge, where the largest obstacle is the diversity and complexity of research techniques. The vocabulary of basic measurement units in e.g. material science is universally understood and accepted, but the research context is rapidly evolving and hard to track. Some of these obstacles can be ameliorated by incorporating contextual metadata at the moment of data creation by automatic means.

The proposed uniform data acquisition infrastructure supports a modern electronic lab book and augments the datasets with the project context and the author identification information. Uploading the data to the repository immediately after the data is collected makes data ready for collaboration and sharing. The repository project can thus become a FAIR data object on its own merit. Since deployment, researchers enthusiastically using the framework, and we already see 30+ OSF projects generated by the MagLab users.

The metadata can be further expanded with supporting information such as the sample description and the provenance info, measurement instrumentation identifiers, probe schematic data and design files, the software versions, the user Open Researcher and Contributor ID (ORCID) information, and even the role of each researcher in data collection. Linking researchers' names and ORCIDs with the datasets will help build an incentive system for propelling data sharing by better crediting researchers for datasets creation [26].

### 3.6 Conclusion

We showcase a FAIR-ready data management framework which non-disruptively integrates existing practices at the user facility. Using today's advances in information technology, we were able to automate data management and reduce the FAIR-compliance burden, making FAIR attractive and beneficial to every

researcher. We open-sourced our data acquisition and metadata hub software, along with APIs to synchronize data to OSF, so other laboratories can build upon the ideas presented here. We believe that our project will inspire other experimental sciences laboratories to carve the pathway to the FAIR-ready data.

## References


1. Actor framework. https://labviewwiki.org/wiki/Actor_Framework. Accessed 30 Apr 2022
2. Bailey, C.B., Balakirev, F.F., Balakireva, L.L.: Closing the gap between FAIR data repositories and hierarchical data formats. Code4Lib (2021)
3. Bray, T.: The JavaScript Object Notation (JSON) data interchange format (2017). https://datatracker.ietf.org/doc/html/rfc8259
4. Cain, B., Klein, M., Finnell, J.: Nucleus - deploying research data management infrastructure at the Los Alamos national laboratory. In: 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp. 396–397 (2019). https://doi.org/10.1109/JCDL.2019.00087
5. The research infrastructure for language as social and cultural data. http://www.clarin.eu/. Accessed 30 Apr 2022
6. Data repository guidance. https://www.nature.com/sdata/policies/repositories. Accessed 30 Apr 2022
7. Infrastructure for the European network for earth system modelling. https://is.enes.org/. Accessed 30 Apr 2022
8. excel. https://www.microsoft.com/en-us/microsoft-365/excel
9. Finnell, J., Klein, M., Cain, B.J.: Nucleus: a pilot project. https://arxiv.org/ftp/arxiv/papers/1705/1705.07862.pdf. Accessed 30 Apr 2022
10. General repository comparison. https://doi.org/10.5281/zenodo.3946720. Accessed 30 Apr 2022
11. Hdf5 file format description and documentation from HDF group. https://portal.hdfgroup.org/display/HDF5/HDF5. Accessed 30 Apr 2022
12. High magnetic field science toolset. https://github.com/ffb-LANL/High-Magnetic-Field-Science-Toolset. Accessed 30 Apr 2022
13. Higman, R., Bangert, D., Jones, S.: Three camps, one destination: the intersections of research data management, FAIR and open. Insights **32**(1) (2019)
14. IGOR. https://en.wikipedia.org/wiki/IGOR_Pro
15. Labview. https://www.ni.com/en-us/shop/labview.html. Accessed 30 Apr 2022
16. Lin, D., et al.: The trust principles for digital repositories. Sci. Data **7**(1), 144 (2020). https://doi.org/10.1038/s41597-020-0486-7
17. Maglab. https://nationalmaglab.org/. Accessed 30 Apr 2022
18. Maglab calendar. https://users.magnet.fsu.edu/Experiments/Calendar.aspx. Accessed 30 Apr 2022
19. Maglab metadata hub. https://github.com/luda171/Maglab-Metadata-Hub. Accessed 30 Apr 2022
20. MATLAB. https://www.mathworks.com/products/matlab.html
21. The novel materials discovery (nomad) centre of excellence. https://nomad-coe.eu/. Accessed 30 Apr 2022
22. numpy. https://numpy.org/
23. Open science framework. https://osf.io/. Accessed 30 Apr 2022
24. Origin. https://www.originlab.com/

25. Osf apiv2 documentation. https://developer.osf.io/. Accessed 30 Apr 2022
26. Pierce, H.H., Dev, A., Statham, E., Bierer, B.E.: Credit data generators for data reuse (2019). https://www.nature.com/articles/d41586-019-01715-4, https://doi.org/10.1038/d41586-019-01715-4
27. Pulsed field facility. https://nationalmaglab.org/user-facilities/pulsed-field-facility. Accessed 30 Apr 2022
28. National Academies of Sciences, E., Medicine: open science by design: realizing a vision for 21st century research. The National Academies Press, Washington, DC (2018). https://doi.org/10.17226/25116, https://nap.nationalacademies.org/catalog/25116/open-science-by-design-realizing-a-vision-for-21st-century
29. Solle, D.: Be FAIR to your data. Anal. Bioanal. Chem. **412**(17), 3961–3965 (2020). https://doi.org/10.1007/s00216-020-02526-7
30. Stall, S., et al.: Make scientific data fair. https://doi.org/10.1038/d41586-019-01720-7. Accessed 30 Apr 2022
31. Stvilia, B., et al.: Research project tasks, data, and perceptions of data quality in a condensed matter physics community. J. Am. Soc. Inf. Sci. **66**, 246–263 (2015)
32. Tdms file format description and documentation from national instruments. https://www.ni.com/en-us/support/documentation/supplemental/06/the-ni-tdms-file-format.html. Accessed 30 Apr 2022
33. The OAuth 2.0 authorization framework. https://datatracker.ietf.org/doc/html/rfc6749. Accessed 30 Apr 2022
34. Turning fair into reality. https://op.europa.eu/en/publication-detail/-/publication/7769a148-f1f6-11e8-9982-01aa75ed71a1/language-en. Accessed 30 Apr 2022
35. Wilkinson, M.D., et al.: The fair guiding principles for scientific data management and stewardship. Sci. Data **3**(1), 160018 (2016). https://doi.org/10.1038/sdata.2016.18