CI Compass
TECH NOTES
ci-compass.org

Making the Major Facilities Data Lifecycle FAIR

Charles Vardeman
Date Published: January 25, 2022

What is FAIR data?

The notion of the four foundational principles for "data" — Findability, Accessibility, Interoperability, and Reusability or "FAIR" — was proposed by Wilkinson et al. in *"The FAIR Guiding Principles for scientific data management and stewardship"* **[1]** and envisages a set of first principles for research communities with respect to the management and curation of scientific data. These principles were created from the point of view that data should be *structured* in a way that the data itself is "smart data" which can be queried for information relative to the four FAIR principles. That is, given the "4 Vs" of big data of Volume, Variety, Veracity and Velocity, it is very difficult for humans to manage data without machine based assistance. The FAIR principles take this need into account for machine actionability to assist humans in understanding and managing data assets. Given the rise of interest for the application of machine learning based surrogates as scientific models, the importance of "AI-ready" structured data will likely become prevalent in the coming years. The FAIR data principles give a foundation for machine learning, and in particular, *Knowledge Informed Machine Learning* **[2]**, that integrates broader knowledge and context into the machine learning process. Specific attributes for each FAIR principle are contained in **Table 1** and require implementation relative to a specific scientific community through community based recommendations.

Table 1: The FAIR Guiding Principles
To be Findable: F1. (meta)data are assigned a globally unique and persistent identifier F2. data are described with rich metadata (defined by R1 below) F3. metadata clearly and explicitly include the identifier of the data it describes F4. (meta)data are registered or indexed in a searchable resource

This project is supported by the National Science Foundation Office of Advanced Cyberinfrastructure in the Directorate for Computer Information Science under Grant #2127548.

To be Accessible:

A1. (meta)data are retrievable by their identifier using a standardized communications protocol

A1.1 the protocol is open, free, and universally implementable

A1.2 the protocol allows for an authentication and authorization procedure, where necessary

A2. metadata are accessible, even when the data are no longer available

To be Interoperable:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (meta)data use vocabularies that follow FAIR principles

I3. (meta)data include qualified references to other (meta)data

To be Resuable:

R1. meta(data) are richly described with a plurality of accurate and relevant attributes

R1.1. (meta)data are released with a clear and accessible data usage license

R1.2. (meta)data are associated with detailed provenance

R1.3. (meta)data meet domain-relevant community standards

Who is adopting FAIR data principles?

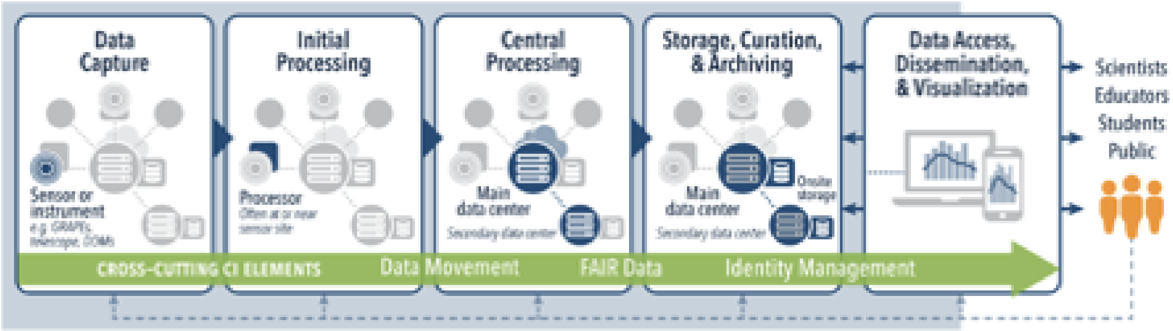
Since the publication of the original Wilkinson et al. paper in 2016, organizations such as “GO FAIR[1]” have been created to encourage adoption by scientific communities and provide forums for scientific communities of practice to tailor recommendations for their specific domain. As of this report, 25 “implementation networks” are actively engaged in defining what the FAIR principles mean to their communities and creating guidelines specific to that communities needs. Wilkinson’s paper was motivated by the life sciences community of practice, one of the largest early adopters of the principles and have led the sciences in active adoption. Commercial entities have also recognized the business value in FAIR data principles and have adopted recommendations from the guidelines in their data infrastructure. The Pistoia Alliance,[2] created by the 2 pharmaceutical industry, has created “The FAIR Toolkit for the Life Science Industry”[3] providing guidance, use 3 cases and examples to motivate industry wide adoption of FAIR principles. Other industry discussions about FAIR principles have revolved around enterprise metadata management best practices. For example, Linked-In sponsored

[1] Go FAIR (https://www.go-fair.org/)

[2] Pistoia Alliance (https://www.pistoiaalliance.org)

[3] Pistoia Alliance FAIR Toolkit (https://fairtoolkit.pistoiaalliance.org)

This project is supported by the National Science Foundation Office of Advanced Cyberinfrastructure in the Directorate for Computer Information Science under Grant #2127548.



“Metadata Day 2020”[4] using an 4 “unconference” style workshop to facilitate discussion about best practices for metadata in commercial enterprises including adoption of FAIR principles[5] in broader enterprise data 5 strategies.

How does FAIR relate to the data lifecycle?

FAIR principles can be seen as a “cross-cutting” cyberinfrastructure element that has broad application for the entire data lifecycle **[3]** model for NSF large facilities. This means that there is a value proposition to major facilities to utilize FAIR during each stage of the data lifecycle, not just in the final **data access and dissemination stage** of the data lifecycle. During **data capture**, **initial processing** and **central processing stages**, **findability** and **accessibility** of the data artifacts are, by definition, accounted for in data movements and workflows associated with the lifecycle. The use of dereferencable persistent identifiers (F1 and A1) can provide additional value for tracing digital objects though the data lifecycle process. For example, AstraZeneca has created a formal identifier policy based on URI Templates where the URI structure aids in data source identification.[6] AstraZeneca’s URI 6 scheme maps data to clinical trial study identifiers throughout the study lifecycle. In this way, it is easy to **access** information about a study phase and directly connect to data artifacts associated with the study.

Additionally, the **findable** and **accessible** principles also apply to metadata associated with the data artifacts. The use of FAIR metadata catalogs that capture **data lineage** through a digital objects lifecycle enhances the data value proposition.[7] Metadata accessibility propagates into **interoperability** and **reuse** through using FAIR vocabularies **[6-7]** and ontologies as the basis for exposing metadata objects. FAIR vocabularies are themselves FAIR data objects that additionally link to and reuse other vocabularies to give broader context to machine agents.

[4] Metadata Day 2020 (https://metadataday2020.splashthat.com)

[5] Metaspeak 2020 Meet up (https://youtu.be/LS2LxEsj-94?t=3491)

[6] Adoption and Impact of an Identifier Policy – AstraZeneca (https://fairtoolkit.pistoiaalliance.org/use-cases/adoption-and-impact-ofan-identifier-policy-astrazeneca/)

[7] The importance of a FAIR Data Strategy in enhancing the Data Value Lifecycle (https://www.thehyve.nl/articles/fair-datastrategy-for-data-value-lifecycle)

This project is supported by the National Science Foundation Office of Advanced Cyberinfrastructure in the Directorate for Computer Information Science under Grant #2127548.

What does it mean for NSF Major Facilities?

With the recent success of deep learning based scientific models, such as the corporately developed Deepmind AlphaFold 2 protein folding model,[8] over traditional physics based computational solutions there is a movement toward adoption of deep learning methodology by many science disciplines. Deep learning requires large volumes of high-quality data, and NSF Major Facilities (MFs) are one source of such data, so *the use of FAIR principles in the data lifecycle will become increasingly important*, and of high value, to researchers utilizing MF resources. Likewise, there is a massive, synergistic, revolution happening with the adoption of deep learning technologies in the private sector that is driving development of components, workflows and pipelines that can be readily adopted into the data lifecycle cyberinfrastructure. The increased interest in open source, cloud-based, FAIR data technologies by both industry and academia can be leveraged thus providing potential partnerships that could lead to more sustainable pathways for FAIR cyberinfrastructure environments.

The speed with which this revolution is happening provides significant challenges to MFs' implementation of FAIR principles in practice. Through industry surveys, Gartner has identified that many of the core areas of expertise relating to FAIR data such as ontologies, knowledge graphs and operationalizing AI initiatives (MLOps, ModelOps) are of great interest to commercial organizations.[9] In the near 9 term, commercial demand for these skills will likely outpace the availability of personnel with appropriate training. However, commercial interest in these technologies should in the long term drive interest in personnel to develop expertise in these technology areas creating a more sustainable workforce. Many synergistic opportunities exist for workforce development in FAIR data and AI operations that could help in recruiting talent to MF operations.

One of the core **missions** for **CI Compass** is to **lower barriers** for MFs to operationalize FAIR data best practices throughout the data lifecycle. This includes providing expertise in FAIR vocabularies, ontologies and metadata as well as cyberinfrastructure components for FAIR data implementation. CI Compass aims to help build a community of practice around FAIR principles for the MFs as well as connecting domain specific communities of practice organized by groups such as the Earth Science Information Partners (ESIP), International Virtual Observatory Alliance (IVOA), and the Research Data Alliance (RDA) that are building the domain specific FAIR tools needed

[8] Putting the power of AlphaFold into the world’s hands. https://deepmind.com/blog/article/putting-the-power-of-alphafold-into-the-worldshands

[9] The 4 Trends That Prevail on the Gartner Hype Cycle for AI, 2021. https://www.gartner.com/en/articles/the-4-trends-that-prevail-on-the-gartner-hype-cycle-for-ai-2021

This project is supported by the National Science Foundation Office of Advanced Cyberinfrastructure in the Directorate for Computer Information Science under Grant #2127548.

by practitioners. CI Compass seeks to assist Major Facilities with the challenges of becoming **AI-ready** for future research and engineering needs.

References

1. Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). DOI: https://doi.org/10.1038/sdata.2016.18

2. L. von Rueden et al., "Informed Machine Learning - A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems," in IEEE Transactions on Knowledge and Data Engineering, DOI: https://doi.org/10.1109/TKDE.2021.3079836.

3. Laura Christopherson, Anirban Mandal, Erik Scott, and Ilya Baldin. 2020. Toward a Data Lifecycle Model for NSF Large Facilities. In Practice and Experience in Advanced Research Computing (PEARC '20). Association for Computing Machinery, New York, NY, USA, 168–175. DOI: https://doi.org/10.1145/3311790.3 396636

4. Ewa Deelman, Anirban Mandal, Angela P Murillo, Jarek Nabrzyski, Valerio Pascucci, Robert Ricci, Ilya Baldin, Susan Sons, Laura Christopherson, Charles Vardeman, Rafael, Ferreira da Silva, Jane Wyngaard, Steve, Petruzza, Mats Rynge, Karan Vahi, Wendy Whitcup, Josh Drake, and Erik Scott. 2021. Blueprint: Cyberinfrastructure Center of Excellence. DOI: https://doi.org/10.5281/zenodo.4587866

5. European Commission, Directorate-General for Research and Innovation, Turning FAIR into reality : final report and action plan from the European Commission expert group on FAIR data, Publications Office, 2018, https://data.europa.eu/doi/10.2777/54599

6. Hugo, Wim, Le Franc, Yann, Coen, Gerard, Parland-von Essen, Jessica, & Bonino, Luiz. (2020). D2.5 FAIR Semantics Recommendations Second Iteration (1.0). Zenodo. DOI: https://doi.org/10.5281/zenodo.5362010

7. Janowicz, K., Hitzler, P., Adams, B., Kolas, D., & Vardeman, C. (2014). Five Stars of Linked Data Vocabulary Use. Semantic Web, 1 Jan. 2014 : 173 – 176. DOI: https://doi.org/10.3233/SW-140135

This project is supported by the National Science Foundation Office of Advanced Cyberinfrastructure in the Directorate for Computer Information Science under Grant #2127548.