IJDC | *Conference Paper*

FAIR Data Implementation at Large-Scale Data Facilities in the USA

Angela P. Murillo
Indiana University-Indianapolis

Don Brower
University of Notre Dame

Jose Cordova
Indiana University-Indianapolis

Sonja Crnevic
Indiana University-Bloomington

Abstract

This paper examines current FAIR data implementation at large-scale data facilities in the USA. A 15-question survey was distributed to facility personnel to explore the current status of FAIR implementation, including progress, barriers, non-barriers, and data management practices. As FAIR implementation considerably increases trust and transparency, this research facilitates understanding the current status and needs for robust FAIR implementation.

Submitted 9 February 2024 ~ *Accepted* 22 February 2024

Correspondence should be addressed to Angela Murillo. Email: apmurill@iu.edu

This paper was presented at the International Digital Curation Conference IDCC24, 19-21 February 2024

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

http://doi.org/10.2218/ijdc.v19i1.924
DOI: 10.2218/ijdc.v19i1.924

Introduction

This paper aims to describe current FAIR data implementation at large-scale data facilities in the USA. For this purpose, a survey was distributed to facility personnel asking participants about their familiarity with FAIR principles (Wilkinson et al., 2016) progress toward implementing these, barriers and incentives to doing so, the resulting value, and how to improve implementation. Participants were also asked about their facilities' data management practices and technologies they use to support data management. Survey participants included scientists, researchers, developers, system administrators, data specialists, and repository managers from large-scale data facilities. As FAIR implementation aims to improve the trustworthiness and transparency of research data, this paper expands understanding of the status and needs for enacting FAIR implementation.

The emphasis on making US federally funded research data publicly available can be traced back to the 2013 Director of the Office of Science and Technology Policy (OSTP) memo (Holdren, 2013). In 2018, the US Congress passed the OPEN Government Data Act, which mandates that data assets belonging to federal agencies be machine-readable. Building on this, the OSTP updated its policy in August 2022 and re-emphasized the need for free, immediate, and equitable public access to federally funded data, promoting transparency and accessibility to advance scientific progress and innovation (Nelson, 2022).

How data are created, managed, and maintained can considerable impact their accessibility, reusability, and machine-readiness. Due to the need to create guidance for ensuring machine-ready research data, diverse stakeholders came together to design the FAIR data principles, which act as data management guidelines (Wilkinson et al., 2016). Implementing FAIR principles provides many advantages to researchers and research facilities, and improves the machine-readiness of research data for the greater scientific community (Berka et al., 2023).

Due to the importance of supporting large-scale data management, CI Compass was established as the National Science Foundation (NSF) Cyberinfrastructure (CI) Center of Excellence, charged with providing expertise and support to CI practitioners and building a CI ecosystem, with a particular focus on the NSF Major Facilities (MFs). MFs represent the largest research CI investments made by the NSF, span a wide-range of scientific disciplines, and rely on complex CI to support the MFs science missions (CI Compass, 2023b). There are currently approximately 20 MFs[1] that NSF supports. CI Compass focuses on providing support for the full data lifecycle (i.e., data capture, processing, curation, access and dissemination), including cross-cutting elements such as FAIR implementation (See Figure 1). As MFs are a significant contributor to the creation of scientific data, it is important to understand current FAIR data practices at MFs and similar large-scale data facilities.

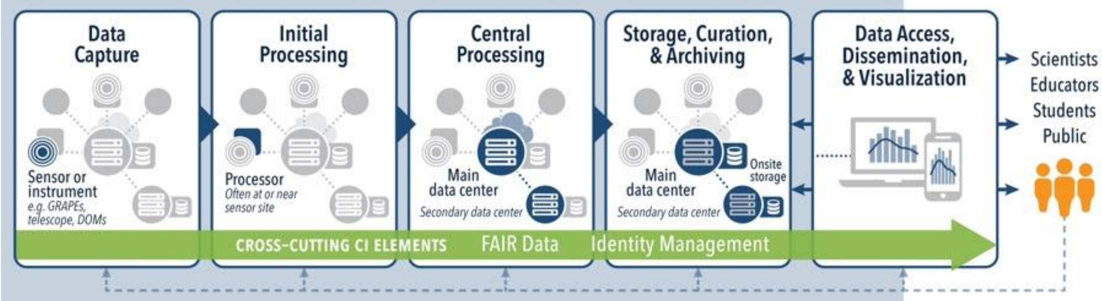

Figure 1. CI Compass Data Lifecycle Model (CI Compass, 2023a)[2]

[1] https://www.nsf.gov/bfa/lfo/docs/major-facilities-list.pdf

[2] CI Compass refers to the NSF Center of Excellence for Navigating the Major Facilities Data Lifecycle. DOMs (digital optical modules) refer to sensors at the IceCube Neutrino Observatory. GRAPEs refer to

In the Spring of 2022, CI Compass hosted the Cyberinfrastructure for Major Facilities (CI4MF) workshop. Workshop attendees included scientists, researchers, and data managers from MFs, mid-scale facilities, and the greater scientific community. Workshop attendees were primarily from the United States, however, MFs are located all over the world.[3] The workshop consisted of panels and discussions of best practices, opportunities, and potential solutions to challenges related to large-scale scientific data management. Based on a pre-workshop survey that called for topics, a session was devoted to FAIR implementation. This workshop session determined the need to further investigate FAIR implementation in large-scale data facilities to explore the current status and best practices (Deelman et al., 2022).

In response to the workshop session, a CI Compass working group was created to share ideas and approaches toward implementing FAIR. The working group consisted of CI Compass personnel, MF personnel, and CI professionals, which include scientists, data managers, and others with expertise in large-scale data management. The working group's objectives are to understand FAIR implementation through researching FAIR data practices and implementation, disseminating research on FAIR data, and organizing guest speakers related to FAIR, particularly in the context of large-scale scientific data cyberinfrastructure.

To explore the current landscape of FAIR implementation, the working group developed a survey to understand the current status, approaches, and best practices. The remainder of this paper provides a review of related literature, survey development and distribution, findings, discussion, and conclusion.

Related Literature

The FAIR Data Principles were introduced in 2016 by a diverse set of stakeholders from academia, industry, funding agencies, and scholarly publishers came together to provide a set of measurable principles to improve the infrastructure supporting the reuse of research data. These measurable principles are to make data findable, accessible, interoperable, and reusable and specifically focus on enhancing the machine-readability of data (Wilkinson et al., 2016). Since 2016, there has been much research and scholarship regarding FAIR data, including application guidance, discipline-specific implementation, tool development, and curriculum development.

There are many incentives to adopting FAIR principles, as making data machine-actionable allows computational systems to interact with data with little to no human effort. Implementing FAIR impacts various stakeholders, including data creators, data reusers, data repository managers, and data infrastructure professionals, as well as funding agencies (Wilkinson et al., 2016). Furthermore, sharing data that complies with FAIR principles serves the scientific community because it facilitates knowledge discovery and increases the possibility of data reuse of data and potential collaboration (Columbia University Irving Medical Center, 2022).

Scholarship regarding FAIR implementation has been widespread and diverse with regard to scientific domains (Van Reisen, Stokmans, Basajja, et al., 2020). There is existing research on its application to materials science (Scheffler et al., 2022), marine science (Tanhua et al., 2019), biopharmaceutics (Wise et al., 2019), health research (Sinaci et al., 2020), radiotherapy (Kalendralis et al., 2021), geoscience (Lannom et al., 2020), natural sciences (Bishop et al., 2019; Bishop & Collier, 2022), agricultural sustainability (Ali & Dahlhaus, 2022; Devare et al., 2021), among many others.

There has been scholarship to provide additional FAIR guidance for data repositories, data infrastructures, and data citation guidelines (Groth et al., 2020; Neumann, 2022). Tools have been developed to help implement FAIR, such as the FAIR Data Point metadata repository

'Grouped Remote Analog Peripheral Equipment' that is used at the NSF National Ecological Observatory Network (NEON) to send sensor measurements digitally from all NEON sites to NEON headquarters. In this figures, DOMs and GRAPEs are provided as examples of the various ways that data are captured at Major Facilities.

[3] For a list of NSF Major Facilities see here: https://www.nsf.gov/bfa/rio/major-facilities-list

(Benhamed et al., 2023). Researchers have explored how ontology matching may be useful for implementing FAIR (Van Damme et al., 2022) and have compared the FAIR principles with other assessment approaches, such as the CoreTrustSeal Trustworthy Data Repositories framework[4] and the Data Stewardship Maturity Matrix[5] (Peng et al., 2022). Additionally, curricula have been developed to help students and researchers understand and implement FAIR principles, including self-guided courses (Oladipo et al., 2022), 'Bring Your Own Data' (BYOD) workshops (Jacobsen et al., 2020), and FAIR competency development guides (Demchenko & Stoy, 2021).

Nevertheless, FAIR implementation has yet to become widespread in the US. As of 2019, most of the implementation had been in Europe (Van Reisen, Stokmans, Mawere, et al., 2020). Challenges and obstacles at the individual research level include the lack of preparation and support, unfamiliarity with FAIR, perceived lack of need, and fear that implementation will incur costs and reduce budgets. Challenges at the institutional level include the lack of expertise and skillset required for implementation, as tools, software, and standards for implementing FAIR are still under development (Bloemers & Montesanti, 2020).

As large-scale data facilities are a significant contributor to capturing, creating, and disseminating data for the global scientific community, it is vital to understand the current status of FAIR implementation at these facilities. Therefore, it is crucial to identify the optimal approaches for implementing FAIR data management by assessing the needs, opportunities, and implementation gaps at MFs and other large-scale data facilities.

Research Methods

As described previously, the FAIR data working group consists of CI Compass personnel, MF colleagues, and colleagues from the broader scientific community, (e.g., data management professionals and users of research facilities). In Spring 2023, the working group developed a survey to explore FAIR implementation at large-scale data facilities. The survey went through several rounds of iteration, discussion, and modification from feedback provided by researchers at large-scale facilities. Additionally, the survey was piloted for survey flow and question construction. It was approved for human-subject research through the Indiana University Institutional Review Board, IRB #17454.

The 16-question survey (See Appendix) asked participants about their familiarity with FAIR, their progress toward implementation, factors helping or hindering that, as well as about the value of FAIR implementation, and what is needed to improve it. Additionally, participants were asked about the technologies used in their facilities and their data management practices. Finally, participants were asked if they were amenable to an interview to discuss FAIR implementation in further detail.

The survey was administered anonymously online via Qualtrics. Participants were recruited through the NSF Large Facilities Office (LFO) Research Infrastructure Communities of Interest portal. Additionally, the LFO distributed the survey through the Research Infrastructure Webinar and Workshop database. Lastly, CI Compass sent the survey recruitment email to their listserv of contacts. These email distributions targeted personnel from MFs, mid-scale facilities, and other large-scale scientific data facilities. The survey was open from February 20th to March 20th, 2023.

[4] https://www.coretrustseal.org/

[5] https://repository.library.noaa.gov/view/noaa/45294

Findings

Findings: Participant Demographics

The survey received 44 complete responses. The survey participants include personnel from various MFs and other large-scale scientific data facilities. MFs represent 27 (61%), and personnel for other facilities represented 17 (39%) of the participants. There are approximately 20 Major Facilities[6] and 32 Mid-Scale Facilities[7] for a total of 52 Facilities targeted. Of the 52 facilities targeted, 14 participants responded from MFs and 22 responded from other facilities.

Among the MFs, Ocean Observatories Initiative[8] represents 11% of the participants, with National Ecological Observatory Network[9] representing 9%, and the Seismological Facility for the Advancement of Geoscience/Geodetic Facility for the Advancement of Geoscience ((SAGE/GAGE), now the EarthScope Consortium[10]) representing 7%. The IceCube Neutrino Observatory[11], the National Optical-Infrared Astronomy Research Laboratory[12], and the National Hazards Engineering Research Infrastructure[13] each represent 5% of the participants. The remaining facilities have 2% representation each. Participants from other facilities include personnel from the Long-Term Ecological Research Network[14], the San Diego Supercomputing Center[15], the Global Ocean Biogeochemistry Array[16], and various research universities.

Managers and supervisors comprise the largest group at 48.1%, followed by researchers (14.8%), data specialists (11.1%), and repository managers (5.6%). Research programmers, software developers, and system administrators account for 3.7% each. Lastly, the "Other" category accounts for 9.3% of the participants, with participants self-identifying as data curators, data scientists, and data managers.

Findings: FAIR Data Implementation

Most participants are Very Familiar (45.2%) with FAIR, while 21.4% are Extremely Familiar. Moderately familiar participants account for 23.8%, and those who are Slightly Familiar represent 7.1%. Only 2.4% of the respondents are Not Familiar with the FAIR at all.

The survey highlights various FAIR-related activities and their implementation at facilities (see Figure 2). Regarding communicating the importance of FAIR, most facilities have engaged in moderate to considerable efforts (90%), with only a small percentage not addressing FAIR at all (5%). When it comes to providing training on FAIR-related concepts, a majority have offered some level of broad instruction on FAIR concepts (75%). However, in-depth or focused training for FAIR implementation is less common (43%). As for developing facility-wide plans for FAIR implementation, a considerable portion of facilities have made substantial progress (69%), while a smaller group of facilities have had little to no progress (31%). Similarly, the creation of schedules for FAIR data implementation shows mixed results, with some facilities demonstrating a moderate to high level of commitment (64%) and others lagging (36%).

[6] https://www.nsf.gov/bfa/rio/major-facilities-list
[7] https://www.nsf.gov/bfa/rio/midscale-research-infrastructure-list
[8] https://oceanobservatories.org/
[9] https://www.neonscience.org/
[10] https://www.earthscope.org/
[11] https://icecube.wisc.edu/
[12] https://noirlab.edu/public/
[13] https://www.designsafe-ci.org/
[14] https://lternet.edu/
[15] https://sdsc.edu/
[16] https://www.go-bgc.org/

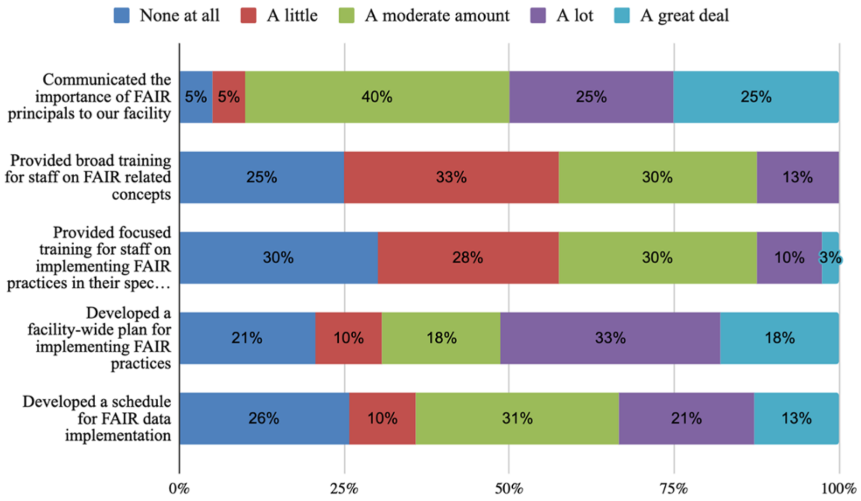

Figure 2. Progress toward FAIR implementation

Participants were also asked through an open-ended question to describe how they assess their progress if there has been work done towards FAIR. These responses indicate that implementing FAIR principles vary across facilities. Some facilities have made progress by adopting FAIR principles in their implementation and data management plans, for example, "*we believe that our implementation plan satisfies FAIR, so FAIR progress is embedded withing our overall program progress and assessment*" and "*my group was adhering to the principles of FAIR before it was called FAIR.*" While others have plans to continue improving their implementation of FAIR, for example, "*we have done well in some aspects...however, one aspect that continues to be worked on is consistent naming standards in line with community standards.*" Some participants face challenges such as lack of resources and funding or resistance from researchers, for example, "*the issue is the lack of person-power and funding to implement FAIR related and a hesitance by researchers to change.*" Additionally, some participants indicated technical challenges, for example, "*we are challenged with a system engineered before FAIR became common practice which means a system we often have to adapt to implement FAIR in less-than-ideal ways.*" Several facilities assess their progress by incorporating FAIR into annual work plans, reviews, and user surveys. Despite challenges, many participants are making incremental progress and striving to make their data more FAIR in the long term.

The survey results illustrate the impact of various hindrances on FAIR implementation (see Figure 3). Most participants experienced mild to moderate hindrance (58%) regarding the lack of tools, while a smaller portion reported considerable obstacles (27%) or no hindrance (15%). Regarding the lack of people who understand FAIR, most participants faced minor to moderate challenges (71%), with fewer indicating no hindrance (26%) or substantial difficulties (3%). Regarding lack of time (prioritizing) FAIR data, many participants experienced considerable to major obstacles (73%), while others reported minimal (12%) to moderate issues (15%). Regarding the lack of money, participants reported a wide range of hindrances, from minimal (18%) to considerable (26%), with a smaller group facing major challenges (18%). Finally, for the "Other" category, participants identified several hindrances to implementing FAIR, including lack of funding agency commitment, unclear mandates, and unwillingness to make data public. These challenges highlight the need for more resources and clearer direction to facilitate FAIR implementation.

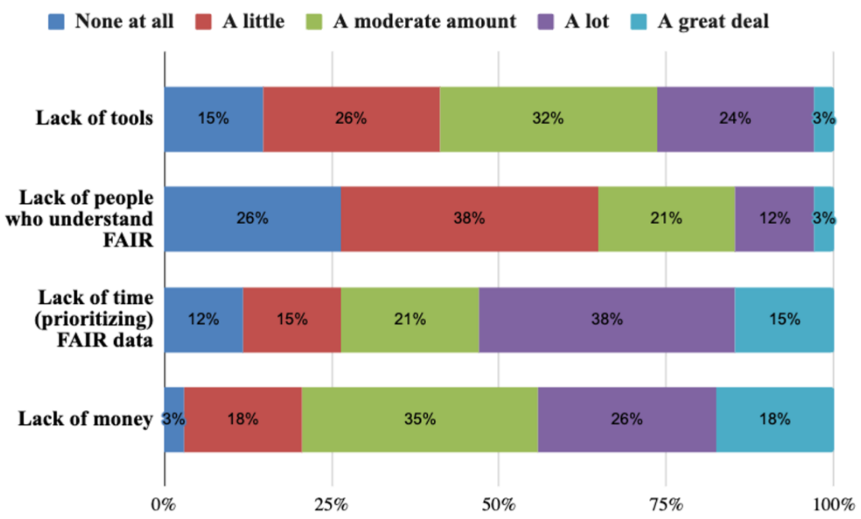

Figure 3. Hindrances for FAIR implementation Progress

The survey results highlight the helpfulness of various factors for in implementing FAIR for their organization (see Figure 4). Technical tools are moderately to very helpful for most participants (91%), with a small portion reporting them as extremely helpful (9%). People who understand FAIR have been predominantly very helpful (15%). Leadership that supports FAIR has shown a considerable impact, with a large percentage of participants finding it very helpful (50%), followed by extremely (32%) and moderately (12%) helpful. Funding has been moderately helpful (33%) for many participants, while also proving very (27%) and extremely helpful (21%) for others. For the "other" category, one participant stated that funding agency commitment to FAIR was extremely helpful for implementation.

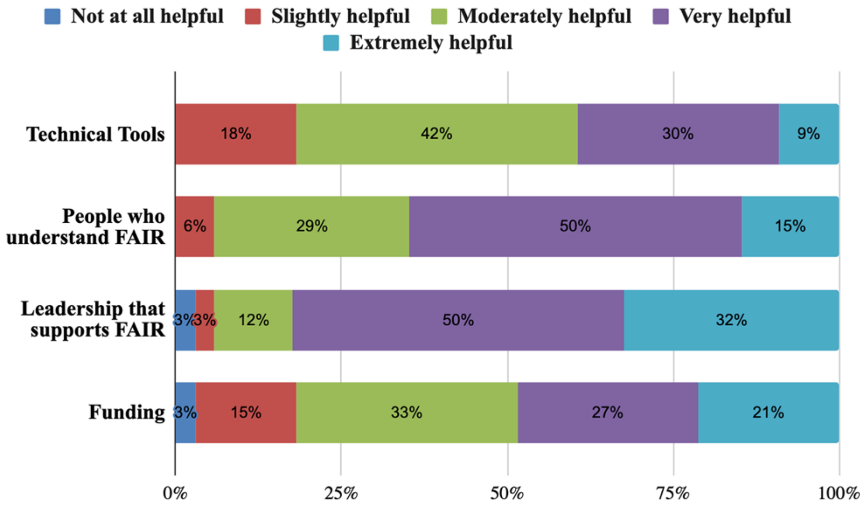

Figure 4. Helpful factors for FAIR implementation (organizationally)

The survey results reveal participants' perspectives on the value of FAIR (see Figure 5). A majority strongly disagree (72%) with the statement "I do not see a value to FAIR," while a few somewhat disagree (3%) or remain neutral (6%). In terms of seeing value in FAIR as a grant requirement, opinions were more diverse, with the majority being neutral (35%) or disagreeing (42%). When considering personal value beyond grant requirements, most participants strongly agreed (64%), with a notable portion somewhat agreeing (18%). The majority of participants strongly agreed (71%) that FAIR practices help their facility, with some also somewhat agreeing (23%). A substantial number of participants strongly agreed (80%) that FAIR provides value to their discipline, with a smaller group somewhat agreeing (14%). Finally, an overwhelming majority strongly agreed (94%) that FAIR provides value to science as a whole, highlighting the participants' overall positive view of FAIR and its impact on scientific research.

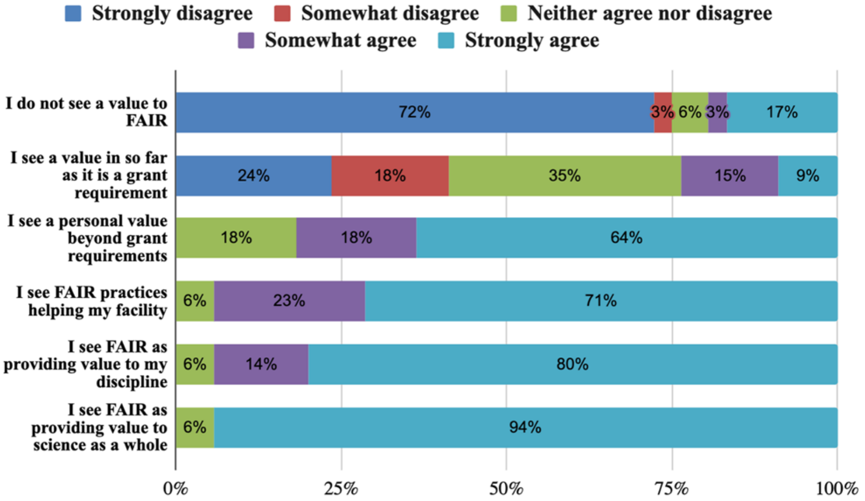

Figure 5. Perceptions of FAIR

The results show the perceived helpfulness of various factors in implementing FAIR principles for the participants individually (see Figure 6). A majority of participants found training on FAIR principles (68%), connecting to implementation networks (79%), and better tooling (81%) to be moderately to extremely helpful. More time (95%)and organizational support (97%) were considered moderately to extremely helpful by more participants. In the "Other" category, participants described how increased funding support, clearer mandates to implement FAIR principles, convincing management on the importance of FAIR, and the availability of tools for tracking citations to data sets would help them become more FAIR.

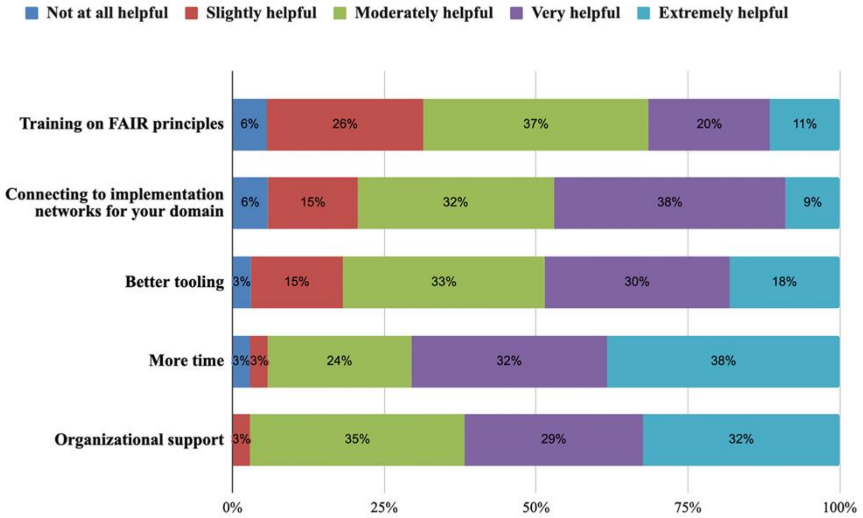

Figure 6. Helpful factors for FAIR implementation (individually)

Findings: Technical Considerations

To gain an understanding of the technologies facilities used to manage their data, Participants were asked to describe the technologies they used for 1) data storage, 2) metadata storage, 3) defined data formats and data models, 4) processing workflows, 5) data and metadata search, and 6) catalogs and websites.

Regarding data storage, the most common technologies mentioned were Filesystem (in different forms) and Cassandra. Other technologies mentioned included S3, CEPH, Lustre, MongoDB, PostgreSQL, MariaDB, MySQL, Google Cloud Storage, NetCDF, NetApp, and multiple file systems. Some facilities also use combinations of these technologies for their data storage needs. Regarding metadata storage, the most common technology mentioned was Postgres. Other technologies mentioned included MySQL, MariaDB, MongoDB, Neo4j, Neptune graph, NetCDF, Oracle, and text files. Some facilities use combinations of these technologies for their metadata storage needs, while others have unique solutions tailored to their specific research domains.

Regarding defined data formats and data models, the most common technology mentioned was NetCDF. Other technologies mentioned included HDF5, Parquet, BagIt, CSV, JSON, XML, TIFF, JPEG, and miniSEED. Some facilities use custom data formats or a combination of different technologies for their data models. A few facilities have not defined specific data formats or use vendor formats and plain-text formats. Regarding processing workflows, custom solutions and code were mentioned several times. Other technologies mentioned included AirFlow, Pachyderm, Condor, Pegasus, Galaxy, DataONE, Frictionless, GitHub, Bitbucket, HTCondor, IceProd, Python, Shell, and web forms. Some facilities use mostly manual processes or proprietary software, while others use customized open-source tools or combinations of different technologies for their processing workflows.

Regarding data and metadata search, the most common technologies mentioned were ElasticSearch and Solr. Other technologies mentioned included API-based database queries, MongoDB, ERDDAP, SQL, PostgreSQL, and web-based tracking tables. Some facilities have no specific solution, while others use custom tools, a variety of tools, or combinations of different technologies for their data and metadata search needs. Regarding catalogs and websites, some facilities use custom solutions or specific platforms such as DKAN, Drupal, DSpace 7, DataOne, AMD/GCMD, Polder, Schema.org, EDI, HUBZero, SeedMeLab, OneSciencePlace, Liferay, React, or Django. Other facilities use popular content management systems like Wordpress or Joomla, while some use web-based tracking tables, text files, or APIs. A few participants were unsure or mentioned that their facility does not have a catalog of data or metadata available.

Additionally, participants were asked various questions regarding data considerations at their facility. A majority of the participants' facilities have datasets with external identifiers (82%) and maintain an index or catalog of available datasets (91%). Most of these indexes or catalogs are available to external users (82%) and offer an API or a way for external computer agents to harvest metadata (75%). Almost half (47%) of the facilities' metadata records are harvested by outside systems. In terms of searching and finding data, half of the participants find their catalog search sufficient, while around 75% need to examine files in the dataset to determine data usability. Notably, 37% of the participants use outside search tools to find their own facility's datasets (see Figure 7).

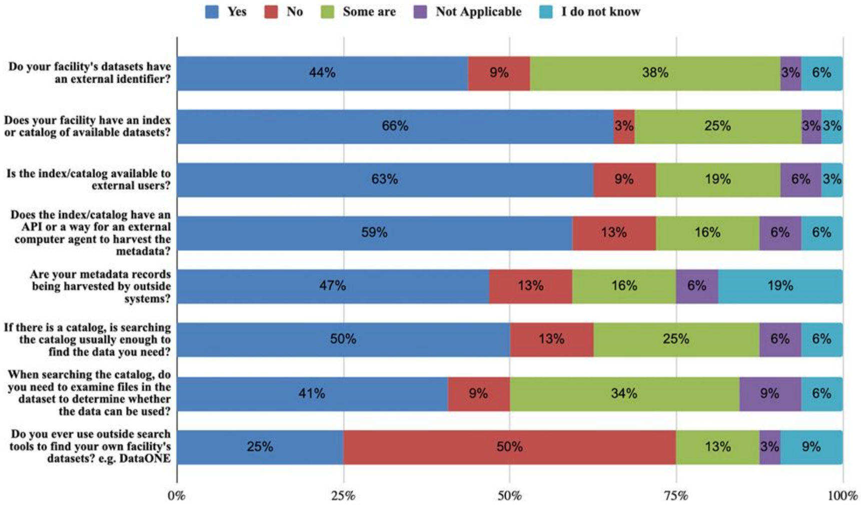

Figure 7. Data Considerations

Findings: Open-Ended Questions

Participants were asked through an open-ended question to describe any approaches for FAIR data implementation that are working particularly well at their institution. Various approaches for FAIR data implementation are working well at different organizations. These approaches include data model representation and tree views for large datasets, CF data compliance with the help of ERDDAP[17] and THREDDS[18], and a commitment to FAIR principles in design and practice.

Some organizations have focused on improving the "findability" and access of their datasets by using the NetCDF format and implementing DOIs and PIDs. Standardizing datasets with interoperable formats like NetCDF/HDF5 and geospatial formats, as well as using standard web service APIs, has also proven effective. Searchable metadata catalogs with Dublin Core discovery-level metadata and domain/data-specific metadata have been employed to further enhance data findability. To ensure data reuse, some organizations have placed an emphasis on tutorials and workshops. Lastly, the implementation of knowledge graphs, ontologies, Schema.org, APIs, flexible metadata, and DOI minting has facilitated better data management. However, some organizations still report cultural barriers, private collaboration data, and a lack of responsibility for data management as challenges to FAIR data implementation.

Participants were also asked through an open-ended question to describe other institutions or organizations they thought were doing a good job with respect to FAIR implementation. Many institutions and organizations are striving to implement FAIR principles, although none are considered perfect. Notable examples include FDSN federated data centers[19], NASA astrophysics data[20], LIGO[21], Rubin Observatory[22], MagLab[23], HEASARC[24], and DataONE[25]. These organizations employ various tools, dedicated staff, thoughtful data rights rules, and uniform data handling practices to promote FAIR data management. Additionally, they explore connections with platforms like GitHub and follow best practices from organizations like Harvard DataVerse[26], FigShare[27], and ICPSR[28] for metadata and social science data management.

A further question asked participants to describe if they saw AI/ML as a driver for FAIR data practices. From participant responses, Artificial Intelligence and Machine Learning (AI/ML) is seen as a driver for FAIR data practices by some participants. The importance of machine readability and the potential for AI and ML to simplify tasks, such as analyzing large datasets, is acknowledged. However, concerns about creating "black box" situations, where researchers do not fully understand AI/ML models or their results, are also raised. While some believe that FAIR practices are critical for ML applications, others are more cautious and think that AI/ML applications should be chosen carefully. The need for detailed and accurate metadata for AI/ML training is emphasized, and some participants highlight potential uses, such as enhancing dataset search, generating semantic knowledge graphs, and linking data to other resources. AI/ML readiness is seen as a driving force for FAIR implementation in certain domain communities.

Lastly, participants were asked through an open-ended question if they had anything else they would like to share with regard to FAIR. Regarding FAIR data implementation, participants emphasize that the human side is the most challenging aspect, and FAIR plans should acknowledge this. Implementing FAIR is seen as an iterative process that requires time, money, training, and tools. The distinction between "closed" and "open" FAIR in the context of large facilities is noted, with the LHC being an example of good "closed" FAIR implementation. The

[17] https://coastwatch.pfeg.noaa.gov/erddap/index.html
[18] https://www.unidata.ucar.edu/software/tds/
[19] https://www.fdsn.org/datacenters/
[20] https://science.data.nasa.gov/astrophysics-data/
[21] https://www.ligo.caltech.edu/
[22] https://rubinobservatory.org/
[23] https://nationalmaglab.org/
[24] https://heasarc.gsfc.nasa.gov/
[25] https://www.dataone.org/
[26] https://dataverse.harvard.edu/
[27] https://figshare.com/
[28] https://www.icpsr.umich.edu/

lack of tools to track citations of datasets using DOIs is seen as a barrier to FAIR data practices, and the importance of human-access aspects is highlighted. Participants also mention the difficulties in achieving reproducibility in large repositories and the challenges posed by the lack of commitment and funding from agencies like NSF. Some participants share their efforts in developing new frameworks to support FAIR practices and applying these practices to their own projects.

Discussion and Conclusions

This study aimed to understand the current status of FAIR implementation at large-scale data facilities by exploring FAIR progress, factors that help and hinder progress, and approaches to FAIR implementation.

Several responses were consistent with our expectations, such as participants considering themselves familiar with FAIR and finding value in the FAIR principles on a broad scale. Participants have generally communicated the importance of FAIR and developed plans for implementation. They identified having leadership who support FAIR as being critical. There was a general view that while technical tools helped support FAIR, the lack of tooling was not the biggest hindrance; instead, time and money were. An interesting finding is that while communication on the importance of FAIR was reported, this communication was not in the form of training at either a broad or detailed level.

Our findings indicated that a majority of participants were either very (45.2%) or extremely familiar (21.4%) with the FAIR principles. However, it was evident that the implementation of these principles varied considerably across different facilities. Some demonstrated considerable progress, whereas others grappled with challenges or exhibited slow progress. Key elements that were deemed useful in implementing FAIR encompassed training, networking with other implementing bodies, improved tooling, organizational support, and increased time allocation.

The major obstacles impeding FAIR implementation included lack of funding, unclear mandates, resistance to making data public, and constraints related to time and financial resources. Despite these challenges, most participants identified technical tools, knowledgeable personnel, leadership support, and funding as beneficial factors aiding FAIR implementation. Moreover, the majority acknowledged the value of FAIR for their individual facilities, for their respective disciplines, and for the scientific community as a whole.

It was also noted that the respondent facilities utilized a diverse range of technologies for data and metadata storage, data formatting, processing workflows, and for data and metadata search, along with catalogs and websites. Interestingly, the organizations exhibited varying strategies in effectively implementing FAIR principles, while some still encountered cultural barriers and challenges. AI/ML was viewed as a motivating factor for adopting FAIR data practices, albeit concerns regarding "black box" situations and the necessity for accurate metadata were raised.

As for the limitations of our study, there is no exact estimate of the target population; therefore, an exact response rate cannot be indicated. Additionally, since this surveyed facility professionals, individual survey participants may have more knowledge of certain aspects of the data management at their facility, but perhaps not knowledge of other aspects. Additionally, survey participation were indicated by the decreasing response rate throughout the survey, possibly reflecting the length of the survey or that FAIR implementation requires many different people and skills, and no one person has full knowledge of all the details. Furthermore, the social desirability bias and the possible loss of nuances in participants' experiences due to the survey methodology were limitations. However, this has been mitigated with follow-up interviews, which have been completed, and data analysis is currently underway and will deepen the understanding of FAIR implementation. The findings from the interview analysis will be made available within the next year.

Lastly, based on the initial findings of the interview analysis, as well as further discussion with the FAIR Data Working Group, the group is conducting additional research, particularly in regard to metrics for measuring FAIR, considerations for instrument persistent identifiers, and concerns with long-term archiving.

Acknowledgements

This project is supported by the National Science Foundation Office of Advanced Cyberinfrastructure in the Directorate for Computer Information Science under Grant #21275488. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Ali, B., & Dahlhaus, P. (2022). The role of FAIR data towards sustainable agricultural performance: A systematic literature review. *Agriculture, 12*(2), 309. https://doi.org/10.3390/agriculture12020309

Benhamed, O., Burger, K., Kaliyaperumal, R., Santos, L., Suchanek, M., Slifka, J., & Wilkinson, M. (2023). The FAIR Data Point: Interfaces and tooling. *Data Intelligence, 5*(1), 184–201. https://doi.org/10.1162/dint_a_00161

Berka, K., Bösl, K., Buono, R. A., D'Anna, F., Fatima, N., Henderson, A., Hjerde, E., Jacquemot-Perbal, M.-C., Kalberg, Y., Martinat, D., & Sarntivijai, S. (2023). *RDMkit.* Machine Actionability - RDMkit by Elixir-Europe.Org. https://rdmkit.elixir-europe.org/machine_actionability.html

Bishop, B. W., & Collier, H. R. (2022). Fitness for use of data: Scientists' heuristics of discovery and reuse behaviour framed by the FAIR data principles. *Information Research: An International Electronic Journal, 27*(3). https://doi.org/10.47989/irpaper942

Bishop, B. W., Hank, C., Webster, J., & Howard, R. (2019). Scientists' data discovery and reuse behavior: (Meta)data fitness for use and the FAIR data principles. *Proceedings of the Association for Information Science and Technology, 56*(1), 21–31. https://doi.org/10.1002/pra2.4

Bloemers, M., & Montesanti, A. (2020). The FAIR funding model: Providing a framework for research funders to drive the transition toward FAIR data management and stewardship Practices. *Data Intelligence, 2*(1–2), 171–180. https://doi.org/10.1162/dint_a_00039

CI Compass. (2023a). *About CI Compass and the Data Lifecycle.* CI Compass. https://ci-compass.org/about/

CI Compass. (2023b). *NSF Major Facilities.* CI Compass. https://ci-compass.org/about/nsf-major-facilities/

Columbia University Irving Medical Center. (2022). *What are the FAIR data principles?* https://library.cumc.columbia.edu/insight/what-are-fair-data-principles

Deelman, E., Baldwin, I., Barnet, S., Berriman, G. B., Brower, D., Casey, R., Chaudhry, S., Christopherson, L., Clark, C., Dobbins, B., Gohsman, M., Kee, K., Livny, M., Mandal, A., Mayani, R., Murillo, A., Nabrzyski, J., Olshansky, A., Pascucci, V., … West, J. (2022). *2022 Cyberinfrastructure for NSF Major Facilities Workshop Report* (Version 1). Zenodo. https://doi.org/10.5281/ZENODO.6643902

Demchenko, Y., & Stoy, L. (2021). Research data management and data stewardship competences in university curriculum. *2021 IEEE Global Engineering Education Conference (EDUCON)*, 1717–1726. https://doi.org/10.1109/EDUCON46332.2021.9453956

Devare, M., Aubert, C., Benites Alfaro, O. E., Perez Masias, I. O., & Laporte, M.-A. (2021). AgroFIMS: A tool to enable digital collection of standards-compliant FAIR data. *Frontiers in Sustainable Food Systems, 5*, 726646. https://doi.org/10.3389/fsufs.2021.726646

Groth, P., Cousijn, H., Clark, T., & Goble, C. (2020). FAIR data reuse – the path through data citation. *Data Intelligence, 2*(1–2), 78–86. https://doi.org/10.1162/dint_a_00030

Holdren, J. P. (2013). *Memorandum for the heads of executive departments and agencies, increasing access to the results of federally funded scientific research.* Office of the President, Office of Science and Technology Policy. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Jacobsen, A., Kaliyaperumal, R., Da Silva Santos, L. O. B., Mons, B., Schultes, E., Roos, M., & Thompson, M. (2020). A Generic workflow for the data FAIRification process. *Data Intelligence, 2*(1–2), 56–65. https://doi.org/10.1162/dint_a_00028

Kalendralis, P., Sloep, M., Van Soest, J., Dekker, A., & Fijten, R. (2021). Making radiotherapy more efficient with FAIR data. *Physica Medica - European Journal of Medical Physics, 82*, 158–162. https://doi.org/10.1016/j.ejmp.2021.01.083

Lannom, L., Koureas, D., & Hardisty, A. R. (2020). FAIR data and services in biodiversity science and geoscience. *Data Intelligence, 2*(1–2), 122–130. https://doi.org/10.1162/dint_a_00034

Nelson, A. (2022). *Memorandum for the heads of executive departments and agencies, ensuring free, immediate, and equitable access to federally funded research.* Office of the President, Office of Science and Technology Policy. https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf

Neumann, J. (2022). FAIR data infrastructure. In S. Beutel & F. Lenk (Eds.), *Smart Biolabs of the Future* (Vol. 182, pp. 195–207). Springer International Publishing. https://doi.org/10.1007/10_2021_193

Oladipo, F., Folorunso, S., Ogundepo, E., Osigwe, O., & Akindele, A. (2022). Curriculum development for FAIR data stewardship. *Data Intelligence, 4*(4), 991–1012. https://doi.org/10.1162/dint_a_00183

Peng, G., Gross, W. S., & Edmunds, R. (2022). Crosswalks among stewardship maturity assessment approaches promoting trustworthy FAIR data and repositories. *Scientific Data, 9*(1), 576. https://doi.org/10.1038/s41597-022-01683-x

Scheffler, M., Aeschlimann, M., Albrecht, M., Bereau, T., Bungartz, H.-J., Felser, C., Greiner, M., Groß, A., Koch, C. T., Kremer, K., Nagel, W. E., Scheidgen, M., Wöll, C., & Draxl, C. (2022). FAIR data enabling new horizons for materials research. *Nature, 604*(7907), 635–642. https://doi.org/10.1038/s41586-022-04501-x

Sinaci, A. A., Núñez-Benjumea, F. J., Gencturk, M., Jauer, M.-L., Deserno, T., Chronaki, C., Cangioli, G., Cavero-Barca, C., Rodríguez-Pérez, J. M., Pérez-Pérez, M. M., Laleci Erturkmen, G. B., Hernández-Pérez, T., Méndez-Rodríguez, E., & Parra-Calderón, C. L. (2020). From raw data to FAIR data: The FAIRification workflow for health research. *Methods of Information in Medicine, 59*(S 01), e21–e32. https://doi.org/10.1055/s-0040-1713684

Tanhua, T., Pouliquen, S., Hausman, J., O'Brien, K., Bricher, P., de Bruin, T., Buck, J., Burger, E., Carval, T., Casey, K., Diggs, S., Giorgetti, A., Glaves, H., Harscoat, V., Kinkade, D., Muelbert, J., Novellino, A., Pfeil, B., Pulsifer, P., ... Zhao, Z. (2019). Ocean FAIR data services. *Frontiers in Marine Science, 6*. https://doi.org/10.3389/fmars.2019.00440

Van Damme, P., Fernández-Breis, J. T., Benis, N., Miñarro-Gimenez, J. A., De Keizer, N. F., & Cornet, R. (2022). Performance assessment of ontology matching systems for FAIR data. *Journal of Biomedical Semantics, 13*(1), 19. https://doi.org/10.1186/s13326-022-00273-5

Van Reisen, M., Stokmans, M., Basajja, M., Ong'ayo, A. O., Kirkpatrick, C., & Mons, B. (2020). Towards the tipping point for FAIR implementation. *Data Intelligence, 2*(1–2), 264–275. https://doi.org/10.1162/dint_a_00049

Van Reisen, M., Stokmans, M., Mawere, M., Basajja, M., Ong'ayo, A. O., Nakazibwe, P., Kirkpatrick, C., & Chindoza, K. (2020). FAIR practices in Africa. *Data Intelligence, 2*(1–2), 246–256. https://doi.org/10.1162/dint_a_00047

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data, 3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Wise, J., de Barron, A. G., Splendiani, A., Balali-Mood, B., Vasant, D., Little, E., Mellino, G., Harrow, I., Smith, I., Taubert, J., van Bochove, K., Romacker, M., Walgemoed, P., Jimenez, R. C., Winnenburg, R., Plasterer, T., Gupta, V., & Hedley, V. (2019). Implementation and relevance of FAIR data principles in biopharmaceutical R&D. *Drug Discovery Today, 24*(4), 933–938. https://doi.org/10.1016/j.drudis.2019.01.008

Appendix: Survey Questions

Q1. What is your primary job role?

- Researcher
- Research Programmer
- Software Developer
- System Administrator
- Manager or Supervisor
- Data Specialist
- Repository Manager
- Other [Please specify]

Q2. What Major Facility or other organization do you work for?

- List + open-ended

Q3. How familiar are you with the FAIR principles?

- Not familiar at all
- Slightly familiar
- Moderately familiar
- Very familiar
- Extremely familiar

Q4. What progress has your facility made toward making research data more FAIR? (1 = None at all; 2= A Little; 3 = A Moderate Amount; 4 = A Lot; 5 = A Great Deal)

- Communicated the importance of FAIR principals to our facility
- Provided broad training for staff on FAIR related concepts
- Provided focused training for staff on implementing FAIR practices in their specific work
- Developed a facility-wide plan for implementing FAIR practices
- Developed a schedule for FAIR data implementation
- Other [Please specify]

Q5. If there has been work done toward FAIR, how do you assess your progress?

- Open-ended

Q6. How much of a hindrance have the following been in regards to implementing FAIR at your Facility?
(1 = None at all; 2= A Little; 3 = A Moderate Amount; 4 = A Lot; 5 = A Great Deal)

- Lack of tools
- Lack of people who understand FAIR
- Lack of time (prioritizing) FAIR data
- Lack of money
- Other [Please specify]

Q7. Which of the following have been helpful in regards to FAIR implementation at your organization?
(1 = Not helpful at all; 2= Slightly helpful; 3 = A Moderate Amount; 4 = A Lot; 5 = A Great Deal)

- Technical Tools
- People who understand FAIR
- Leadership that supports FAIR
- Funding
- Other [Please specify]

Q8. Do you see value in applying the FAIR principles to data generated at your facility?
(1 = Strongly disagree; 2= Somewhat disagree; 3 = Neither agree nor disagree; 4 = Somewhat agree; 5 = Strongly agree)

- I do not see a value to FAIR
- I see a value in so far as it is a grant requirement
- I see a personal value beyond grant requirements
- I see FAIR practices helping my facility
- I see FAIR as providing value to my discipline
- I see FAIR as providing value to science as a whole

Q9. What would help you most in becoming more FAIR?
(1 = Not helpful at all; 2= Slightly helpful; 3 = A Moderate Amount; 4 = A Lot; 5 = A Great Deal)

- Training on FAIR principles
- Connecting to implementation networks for your domain
- Better tooling
- More time

- Organizational support
- Other [Please specify]

Q10. To the best of your knowledge what technologies does your facility use for: (Open-ended)

- Data storage (e.g. Filesystem, Cassandra, etc)
- Metadata storage (e.g. Postgres, Neo4j, etc)
- Defined data formats and data models (e.g. Parquet, HF5, BagIt)
- Processing workflows
- Data and metadata search (e.g. Solr, ElasticSearch)
- Catalogs and websites

Q11. To the best of your knowledge:
(1 = Yes; 2= No; 3 = Some are; 4 = Not applicable; 5 = I do not know)

- Do your facility's datasets have an external identifier?
- Does your facility have an index or catalog of available datasets?
- Is the index/catalog available to external users?
- Does the index/catalog have an API or a way for an external computer agent to harvest the metadata?
- Are your metadata records being harvested by outside systems?
- If there is a catalog, is searching the catalog usually enough to find the data you need?
- When searching the catalog, do you need to examine files in the dataset to determine whether the data can be used?
- Do you ever use outside search tools to find your own facility's datasets? e.g. DataONE

Q12. Are there any approaches for FAIR data implementation that are working particularly well at your organization? If so, please describe.

- Open-ended

Q13. Are there any other institutions or organizations that you think are doing a good job with respect to FAIR? Please describe theses approaches.

- Open-ended

Q14. Do you see AI/ML as a driver for FAIR data practices? If so, please describe.

- Open-ended

Q15. Is there anything else you would like to share regarding FAIR data implementation?

- Open-ended

Q16. Would you be willing to talk with us for a 30-45 minute interview regarding inhibitors and facilitators to implementing FAIR at your institution?

- Yes – contact information form
- No