National Science Foundation
WHERE DISCOVERIES BEGIN

Final Report:

FAIR Hackathon Workshop for Mathematical and Physical Sciences Research Communities[1,2]

[1] Editors: Michael Hildreth (Notre Dame), Natalie Meyers (Notre Dame)
[2] Funded by NSF Award NSF-OAC-1839030.

Table of Contents

Overview

The FAIR Hackathon Workshop for Mathematics and the Physical Sciences (MPS) February 27-28, 2019 in Alexandria, Virginia brought together forty-four stakeholders in the physical sciences community to share skills, tools and techniques to FAIRify research data[3, 4, 5]. As one of the first efforts of its kind in the US, the workshop offered participants a way to engage with FAIR principles (Findable, Accessible, Interoperable and Reusable) Data and metrics in the context of a hackathon. The workshop was designed to address issues of public access to data and to provide experience with FAIR tools and relevant hands-on experience for researchers. Existing FAIR tools and infrastructure were introduced. Hands-on hackathon breakout time was devoted to testing FAIR metrics and tools against physical sciences data. The hackathon invited MPS research data management stakeholders to react to the FAIR principles and to jointly consider gaps in the MPS data sharing ecosystem in the context of researcher's actual projects. FAIR Gap analysis was introduced as a way to identify community-specific tools or infrastructure that could dramatically enhance the ability of domain scientists to make their data more FAIR.

Motivation

The datasets created by different branches of science, especially those supported by the National Science Foundation, are incredibly diverse in terms of size, content, and intellectual accessibility. The FAIR Principles can help researchers determine how and what to preserve for consumption by others, and to guide the manner in which the data will be stored and accessed so that reuse is possible. Tools and procedures exist for making data FAIR, but their adoption in many disciplines across research communities in the United States lags behind their European counterparts. There are many reasons for this, including the perception (real or imagined) that the existing tools are too hard or time-consuming to use, or that the available tools do not fit the needs of a given research project. The result is that, all too often, data is not shared, or an inordinate amount of effort is expended in the invention of tools and procedures that only apply to specific use cases. This workshop brought together small groups of investigators sharing similar preservation needs with an opportunity to gather with FAIR experts to try available tools and learn about emerging infrastructure. The overall group was charged to focus discussion on "FAIR" data efforts for chemistry, materials science, and physics to raise the probability that identified common solutions to implementing FAIR in the US can be better recognized and acted upon. This workshop also gave researchers an opportunity to explore how FAIR principles applied directly to data in their own projects, providing unique insights into what their needs might be in order to complete the process of making their data FAIR.

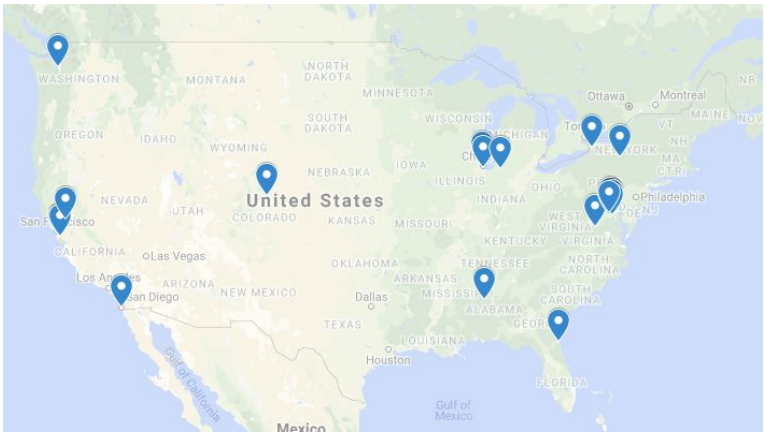
More importantly, as outlined below, the workshop provided an opportunity for physical sciences researchers and data stewards to analyze what is missing to make the pursuit of FAIR data easier for their broader disciplinary communities. Many isolated, uncoordinated data sharing efforts exist under the aegis of several different agencies. Therefore, a diverse group of representatives from across different agencies and federally funded programs (Including NSF, CISE, BRDI, CODATA and NIST)

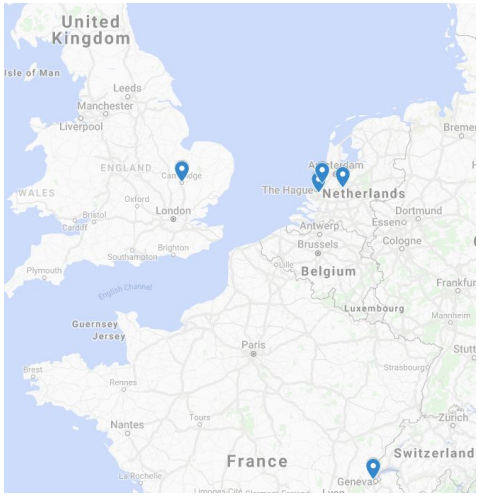
[3] https://mpsfair.crc.nd.edu/
[4] https://osf.io/km8db/
[5] NSF Award OAC-1839030

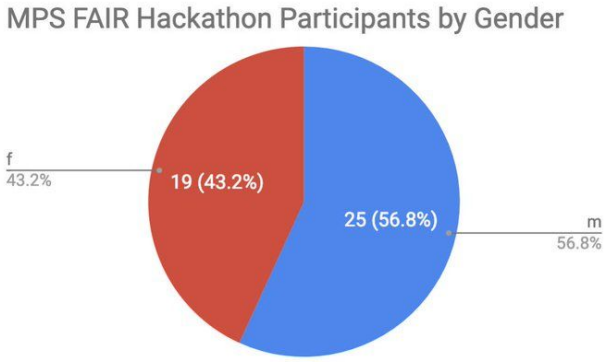
was deliberately recruited alongside participation from ACS, AGU, ACS, NIH/NLM, and USGS). This diverse group of stakeholders served to raise awareness of cross-agency needs among a community of scientists with similar interests and problems. Inter-agency efficiencies in delivering future training and for developing platforms with common aims for FAIR-enabled search and analysis tools can open new areas of interdisciplinary research and cooperation. In the months following this workshop, for example, several workshops auguring the beginning of new efforts in the US were held, including several sessions at the Research Data Alliance Plenary in Philadelphia[6], "Implementing FAIR Data for People and Machines: Impacts and Implications" at the NAS[7] and the Research Data Frameworks (RDAF) workshop[8] at NIST. Clearly, the scope of participation and interest in FAIR principles is expanding rapidly. This report highlights how the unique insights raised by individual researchers can contribute to this global discussion.

Participation

The hackathon convened forty-four participants from across the United States along with FAIR experts, chemistry and physics research counterparts from The Netherlands, the United Kingdom, and CERN/Switzerland.







Workshop participation was 43% female and 57% male. Women and men were at the hackathon in a variety of roles. The fraction of women overall was much larger than is typical of MPS gatherings, especially those revolving around data, and as compared to NCSES, ACS, APS, and AIP society equity numbers for physical sciences.[9,10,11,12]

[6] https://www.rd-alliance.org/rda-13th-plenary-programme
[7] https://www.niso.org/events/2019/09/implementing-fair-data-people-and-machines-impacts-and-implications
[8] https://www.nist.gov/news-events/events/2019/12/research-data-framework-rdaf-workshop
[9] https://ncses.nsf.gov/pubs/nsf19304/digest/field-of-degree-women#physical-sciences (2016, latest year for which NCSES has data readily shared). Women's share of bachelor's, master's, and doctorate degrees in the broad field(s) of physical sciences was 39%, 36%, and 31%, respectively. Physics had the lowest share of women degree recipients, with 19-22% of degrees awarded to women.

FAIR Principles in Detail

Throughout this report, we will refer to the detailed explication of the FAIR principles[13], reproduced here with hyperlinks to more content for convenience:

- FAIR Principles
    - F1: (Meta) data are assigned globally unique and persistent identifiers
    - F2: Data are described with rich metadata
    - F3: Metadata clearly and explicitly include the identifier of the data they describe
    - F4: (Meta)data are registered or indexed in a searchable resource
    - A1: (Meta)data are retrievable by their identifier using a standardised communication protocol
    - A1.1: The protocol is open, free and universally implementable
    - A1.2: The protocol allows for an authentication and authorisation where necessary
    - A2: Metadata should be accessible even when the data is no longer available
    - I1: (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
    - I2: (Meta)data use vocabularies that follow the FAIR principles
    - I3: (Meta)data include qualified references to other (meta)data
    - 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

Pre-Workshop Questionnaire & Participant Responses Indicating Gaps to achieving FAIR

A pre-workshop questionnaire (https://osf.io/e2f3a/) was circulated to gather information about challenges and gaps prior to the event. Participants' responses (https://osf.io/vc9ns/) are shared on the workshop's osf site at: DOI 10.17605/OSF.IO/E2F3A. Some highlights directly related to FAIR gaps are summarized and discussed below.

Participants identified a ***gap between funders' expectations of FAIR readiness and available training***. Prior to the workshop only fifteen participants (35%) responded that their organization, repository provider, or disciplinary society currently offered FAIR-related training materials or training programs.

[10] https://www.acs.org/content/acs/en/membership-and-networks/acs/welcoming/diversity.html (2012, latest year for which ACS has data readily shared)
[11] https://www.aps.org/programs/education/statistics/womenphysics.cfm (2017, latest year for which APS has data readily shared, showing ~20% of physics degrees earned by women)
[12] https://www.aip.org/statistics/reports/women-physics-and-astronomy-2019 (In 2018, 23% of astronomy department faculty members were women while 19% of physics faculty members were women.)
[13] https://www.go-fair.org/fair-principles/

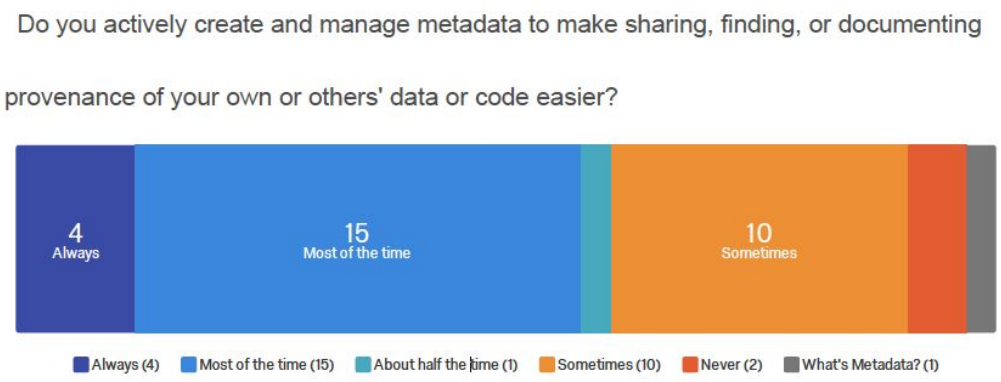
Prior to the workshop 14 participants had data or code to make more FAIR that they were willing to share as examples at the workshop. Five more were unsure if their data was FAIR-ifiable. Ten participants expected to bring no data or code of their own, but had expertise and willingness to help FAIRify others' data/code. Remaining attendees expected to participate in advisory, learning, demonstration, or training roles.

Prior to the workshop only ten participants (22%) noted that they, their funder, or their organization used Data Management Planning Tools. ***The gap between participants' use (and awareness of) data management planning tools was larger than anticipated and therefore prioritized in our workshop programme's hands-on time.***

In a FAIR dataset, F1: (Meta)data are assigned globally unique and persistent identifiers. The majority of participants were familiar with using identifier registration services to assign globally unique identifiers like DOIs that resolve to datasets. Over 23 participating respondents reported familiarity using DOIs and many were able to name a service (like DataCite, CrossRef, or Zenodo) that minted or resolved DOIs for their data, as well as ID services for organizations (ROR), chemicals (InChI), funders and more.

Identifiers exist for researchers as well. Prior to the workshop 26 participants (59%) reported that they had an ORCID. ORCID is a persistent digital identifier for researchers that helps connect research to researchers that supports automatic links between researchers, and their various outputs. After the workshop 40 (91%) of the workshop participants had registered for ORCIDs. This uptick illustrates how quickly ***the gap can be filled for ensuring FAIR F1 compliance on Researcher IDs through awareness campaigns, training/support, and increasing ubiquity of tools that offer ORCID single-sign on and/or ORCID enabled data linking***.

Related to FAIR principle F2: Data are described with rich metadata, the majority of responding participants reported being able to actively create and manage metadata for physical sciences data and code. **We identified little gap for F2 .**



Related to FAIR principle F4: metadata are registered or indexed in a searchable resource, ***there were F4 gaps related to participants' data being searchable and/or in researchers' knowledge about how data/metadata are harvested for indexing by search engines***. Thirteen knew which search engines harvested their (meta)data but an equal number didn't know and some participants (6) acknowledged with certainty that their data was not indexed by search engines). ***The F4 gap might be***

mitigated for physical sciences research data through disciplinary identification of trusted repositories which could then, in turn, ensure disciplinary societies that compliant repositories present (meta) data for search and indexing using standards and protocols that ensure FAIRness.

Similarly, related to FAIR principle A1: (meta) data are retrievable by their identifier using a standardized communications protocol, only eight respondents knew with certainty which protocols controlled access, and the majority of pre-workshop respondents (18) either didn't know what protocols controlled access to their data in the repositories they used, or (5) said their data wasn't finadable via automated search. ***The A1 gap might be mitigated for physical sciences research data through disciplinary identification of trusted repositories*** which could then, in turn ensure disciplinary societies that compliant repositories employ access protocols that ensure FAIRness.

Related to persistence policies and FAIR principle A2: metadata are accessible even when the data are no longer available the majority of pre-workshop respondents (21) didn't know what the persistence policy related to data and metadata in the repositories they used. Three respondents knew with certainty that their repositories had no defined (meta) data persistence policy. Only eight knew what their repository's persistence policy was. ***The A2 gap might be mitigated by funders' requiring deposit to trusted disciplinary repositories***, and the requirement that trusted repositories post their persistence policies and/or make the persistence policy-type query-able or displayed for each (meta) data record if persistence level varies across records or deposits in a given system.

Related to Vocabularies and Ontologies, FAIR principle I1: requires that Metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation (for example the International Union of Pure and Applied Chemistry nomenclature) and FAIR principle I2 requires that metadata use vocabularies that follow FAIR principles. Respondents were evenly split in being able to identify with certainty (16) whether their repository used an ontology(ies) and whether they didn't know (14) and/or knew for certain their repository had none (2) . ***The "I" gaps will be partially mitigated by work being undertaken by the FAIR Convergence Matrix Working Group.***

Related to FAIR principle R1.1: (meta) data are released with a clear and accessible data usage license, we asked MPS workshop participants to identify what licenses they used to indicate permissible data reuse . Twenty-one respondents were able to name their preferred license. Of those, fourteen named Creative Commons, and others mentioned BSD, GPL, MIT and NIST specific licenses. Five acknowledged that they didn't typically assign licenses to their data and code. Five didn't know what their preferred license would be. ***R1 gaps could be mitigated through shared re-usable employer, funder, and disciplinary FIPs and/ or awareness campaigns that specifically included preferred or acceptable resolvable license choices particular to R1.1.***

Related to provenance metadata descriptions and FAIR principle R1.2: (meta)data are associated with detailed provenance, we asked participants what provenance metadata descriptions they were using or supporting in the repositories they maintained. Only three respondents were able to identify a preferred provenance metadata description. Two mentioned NIST's work in this area and another mentioned PROV. Twenty one responded that they "didn't know" and seven responded that they didn't use/supply provenance metadata. ***R1.2 gaps could be mitigated through funder convergence on viable solutions (Like NIST's promises to be) and/or through trusted repository support for re-usable FIPs which declare funder and disciplinary choices for implementing R1.2***

It is clear from the pre-workshop questionnaire that even researchers familiar with FAIR principles have many gaps in their detailed knowledge of how to make data FAIR. Many of these gaps fall into categories such as ontology descriptions and interoperability protocols where one would not necessarily expect domaine researchers to possess expertise, yet these ingredients are essential to achieving the full potential of shared data. This suggests that the creation of a FAIR data ecosystem will require an interdisciplinary partnership of domain researchers, data specialists, and archive providers, or at least some combination of these roles, because the range of expertise required is too broad for an individual, even for one accomplished at their particular role.

Workshop Description

The workshop was structured as an introduction to various aspects of the FAIR principles, the process of FAIR-ifying data, and tools to validate and to quantify progress in this process. The workshop programme and presentations are available online[14]. Because this was the first workshop of its kind in the US and many of the participants were not FAIR experts, a large amount of expository material was included. In the end, this limited the amount of time the participants had to grapple with FAIRifying their own datasets or code, though all had the opportunity to run the FAIR metric tools against their data. Future iterations of workshops like this should aim for more working time with data and code. Below, we outline various components of the workshop and comment on their intended purpose.

Erik Schultes, of GOFAIR, presented on status of FAIR in the US and delivered an overview of efforts world-wide[15] and later Schultes and Albert Mons presented on the GOFAIR Matrix[16] and FAIR Service Provider Consortium[17]. They emphasized the following high level observations:

- FAIR is: Data and services that are findable, accessible, interoperable, re-usable both for machines and for people.
- FAIR Data is a means to an end
- The end goal is NOT FAIR Data but better analytics, more efficiency and impact on research ROI
- The Internet of FAIR Data and Services is the ‘vehicle’
- FAIR tooling supports a data life cycle process
- The FAIRifcation process requires professional services

To acquaint participants with the FAIR principles in detail Ted Habermann of Metadata Gamechangers led the group in a round of the The FAIR metadata game[18]. By the end of the game all participants had collaborated across the metadata lifecycle to create a repository of ten metadata records that included all FAIR elements.

[14] https://osf.io/km8db/
[15] https://osf.io/kr7qp/
[16] https://osf.io/j3xrh/
[17] https://osf.io/sxuc5/
[18] https://osf.io/pj2z5/ (The Metadata Game was developed by Erin Robinson and Ted Habermann).

Disciplinary Presentations

Disciplinary presentations helped MPS FAIR workshop attendees become familiar with FAIRification strategies in health, earth science, chemistry and physics.

Health

Albert Mons presented[19] on training the next generation of enablers of the Internet of FAIR data and Services, describing a Dutch FAIR data stewardship certification and emphasizing that as FAIR data stewardship requires more trainers. There will be an early emphasis on train the trainer events in the EU and the same is recommended for the USA. Mons followed up with inspirational organizational and individual use cases that illustrate the value of discovery based on FAIR integrated interoperable data from an Intra Arterial Trombectomy pilot and Chronic Obstructive Pulmonary Disease (COPD).

Earth Science

Shelley Stall presented on the success of the American Geophysical Union's Enabling FAIR data project[20] where FAIR-aligned earth, space, and environmental science publishers have joined with over 100 signatories to align their policies and establish a similar experience for researchers. Data are no longer placed in the supplemental information; rather data, software, technology are made available through citations that resolve to repository landing pages and availability statements are provided. FAIR-aligned data repositories in turn can add value to research data, provide metadata and landing pages for discoverability, and support researchers with documentation guidance, citation support, and curation. Stall also presented on the PARSEC[21] project which is building new tools for data sharing and reuse through a transnational investigation of socioeconomic impacts of protected areas. PARSEC's tools will improve linkage between data, publications and researchers and promote the incentive credit for open and FAIR data management and preservation for data re-use. Nancy Hoebelheinrich presented on the data stewardship training resources available in the Earth Science Information Partners' (ESIP) data management training clearinghouse (DMTC)[22] and a thoughtful discussion ensued about how the resources could be used to springboard training in physical sciences.

Chemistry

Ian Bruno presented on aspects of FAIR Crystallographic Data[23] and the way a representation of chemistry can be included in a Crystallographic Information Format (CIF). (However, in deposited files this is rarely found.) Then he illustrated how assignment of chemical attributes is required to make crystallographic data findable, interoperable and reusable and followed up with an explication of the enablers of FAIR crystallographic data, including: CIF and dictionaries, as well as standard identifiers and associated infrastructure (DOI, ORCID, InChI . . .). In the case of InCHI alone, this unique identifier has enabled references between the Cambridge Structural Database and PubChem, ChemSpider, the Protein Data Bank, DrugBank, The Pesticides Properties DataBank. Continued growth in InChI facilitated interoperability between systems is anticipated.

---

[19] https://osf.io/ygfp5/
[20] https://osf.io/2ys7r/
[21] https://osf.io/kbxup/
[22] http://dmtclearinghouse.esipfed.org/
[23] https://osf.io/s7nm9/

Stuart Chalk presented on the IUPAC Gold Book: Compendium of Chemical Terminology[24] and its digital evolution, culminating in a demonstration of the new site[25] and an explication of its REST API. He described the future of the Gold Book and how developming an IUPAC chemical ontology will support semantic chemical annotation in compliance with FAIR principle I2. Chalk followed up with a show and tell explicating the role of vocabularies in the context of the Resource Description Framework [26].

Evan Bolton presented on Making Data Interoperable: PubChem Demo/Use case[27] describing the NCBI PubChem (https://pubchem.ncbi.nlm.nih.gov/) resource which serves as an archive of chemical substance information and their biological activities and caters to data scientists through programmatic access to information and machine readable formats. The archive supports many links between large record collections, including: ~245M substances ⇄ ~95M compounds, ~235 M bioactivities ⇄ ~1MBioassays, and ~3M patents ⇄ ~20M compounds. PubChem data sources provide provenance information about what data comes from whom in the form of detailed data downloadable source information. PubChemRDF is comprised of upward of 137B triples. Bolton described how PubChem helps make chemical content findable and leverages WWW and disciplinary chemistry standards. He concluded with a warning that, in spite of this, chemical toolkits don't always behave the same for the same structure and warned that chemical structure information can (irreversibly) change when exchanging between file formats and software packages. He acknowledged that annotating and FAIR-ifying scientific content can be difficult to navigate ad that we must rely on: "Identifiers, licensing/IP, standards,terminologies, normalization, best practices, machine accessibility,scientist education…" all the while realizing that "What you can do today may be different from tomorrow" and that "Everything is a work in-progress." Bolton emphasized that therefore it paramount that "free flow of chemical information makes establishment of best practice, adherence to standards, and scientist education of utmost importance."

Particle Physics

Pamfilos Fokianos presented "Tools to Improve Preservation and Re-Use of Research Results at CERN[28]" including FAIR Ecosystem demonstrations of the INSPIRE High Energy Physics information system, the CERN Document Server, the REANA Reusable Analysis platform and the CERN Analysis Preservation Portal. This work highlights the efforts within the high energy physics community to provide structures for knowledge preservation, linking analysis software, processing structures, datasets, and publications in an information ecosystem. Of particular interest is the REANA infrastructure, which provides a means of preserving and reinstantiating extremely complicated workflows with processing steps encapsulated in unix containers. Subgroups within the large LHC experiments have begun using this system for analysis preservation and are even making the encapsulation of analyses a requirement for publication. The development of suitable vocabularies for classification and searching for these analysis elements is still a work in progress, and, for now, the preserved analyses are only accessible to members of the same experimental collaborations. Thus, achieving full FAIR status for LHC data and analyses is a longer-term project.

---

[24] https://osf.io/n3sk8/
[25] http://dev.goldbook.iupac.org/
[26] https://www.w3.org/RDF/
[27] https://osf.io/pj2z5/
[28] https://osf.io/vq5bs/

FAIR Data Stewardship and Data Management Plans

Because less than a quarter of our participants were familiar with Data Management planning tools we emphasized that during hands-on time during the workshop. NSF's DMP requirement, as stated in NSF 15-052[29] ensures that every proposal submitted to NSF should include a Data Management Plan describing how activities described in the grant proposal will conform to NSF policy on the dissemination and sharing of research results. Participants saw demonstrations of a FAIR specific Data Stewardship Wizard (https://ds-wizard.org/), followed by exposure to a new FAIR tool manager (https://www.fairtoolmanager.com/) in development by PurplePolarBear, a Utrecht technology company in The Netherlands. During hands-on time, they could work with PurplePolar Bear staff to explore their own data in the FAIR tool manager as well as work with workshop coordinators & advisors to contrast features in the Data Stewardship Wizard with existing features available in the Data Management Plan Tool (dmptool.org) which has templates available specific to the requirements of many funding programs in the US, including NSF, NIH, DOE and more. The area of Data Management Planning is evolving rapidly at this point, with many different tools available. A consensus on which directions to emphasize moving forward has yet to emerge.

FAIR Implementation Profiles

At the workshop, FAIR Matrix Working Group member Erik Schultes presented on FAIR implementation profiles (FIPs) and how physical sciences stakeholders can ensure FAIR compliance with I1 and I2 and more through creation of FIPs for disciplinary research in the physical sciences through provision of FIPs authored or vetted by disciplinary societies and/or physical sciences funders like NSF. Subsequent to the MPS hackathon, the Matrix WG has held two development meetings in June 2019 and later in Nov 2019 and published on the *FAIR convergence matrix: optimizing the re-use of existing FAIR-Related resources*[30]. The Matrix has evolved out of the creation of FAIR Implementation Networks[31] that cluster groups of researchers together to share tools, information, and experience in order to FAIRify their data more efficiently. The structures discussed here are an interesting development and provide a useful set of resources to prevent researchers from having to re-invent FAIR-related processes as they move towards creating their own FAIR data.

Fair metrics

MPS FAIR Workshop participants saw presentations about and got to be hands on with tools for Self reporting FAIRness of data sets like the Purple Polar Bear Tool and the FAIR Metric Evaluator. Participants also saw demonstrations of emerging FAIRness tools aimed at repository managers, funders, and research compliance offices that could feature FAIR Reporting, compliance, and endorsement capabilities for the FAIRness of datasets, and the FAIR readiness of repositories.

Participants had feedback time to discuss the potential mis-applications of FAIR metrics and

---

[29] NSF (2015) Today's Data, Tomorrow's Discoveries. https://www.nsf.gov/pubs/2015/nsf15052/nsf15052.pdf

[30] H.P. Sustkova, K.M. Hettne, P. Wittenburg, A. Jacobsen, T. Kuhn, R. Pergl,... & E. Schultes. FAIR convergence matrix: Optimizing the reuse of existing FAIR-related resources. Data Intelligence 2(2020), 158–170. doi: 10.1162/dint_a_00038

[31] https://www.go-fair.org/implementation-networks/

consequences for researchers and repositories. There were concerns expressed regarding how repositories or technologies used in research projects might not be well-FAIRified and how could negatively affect the FAIRness 'score' of researchers. For example, implementing FAIR reporting or metrics along an aspirational continuum could be problematic and fraught with tension for repositories who manage heterogeneous and legacy/pre-FAIR data sets and simultaneously auto-expose FAIRness scores for all. Disciplinary researchers recognized that data curators' interest in tools for measuring and improving data quality, interoperability, and reusability are important but have legitimate concerns about "yet another over-simplistic but reportable way" of providing a metric for tenure and promotion or for documenting comparative departmental "success."

Further FAIR Gaps

Discussions of FAIR Gaps surfaced areas where further training, more developed tools, and clearer FAIR practices would be beneficial for physical sciences researchers and FAIR research broadly. The following needs or gaps were identified during the discussion and subsequent exchanges after the workshop. These can be grouped by overall topic, with no clear order here in terms of importance.

- Tools:
  - There is a general lack of tools integrated within the research process to enable data and results to be made FAIR. Basic ingredients such as provenance-tracking, simple metadata generation, etc., are common to most needs and could be provided with common tools.
  - Basic building blocks of metadata infrastructure, such as a way to describe software, are still under construction or are not widely adopted.
  - There is interest in tools to make the time-consuming process of retrospectively FAIRifying[32] research data and research data repositories more automated.
  - The tools that are available are difficult to find.
- FAIR Metrics:
  - Schema-specific and disciplinary-specific FAIRness indicators are needed.
  - Need to reach understanding/agree on appropriate collection-level FAIRness metrics - because assessing researchers or research project FAIRness at the level of a single shared piece of code or a stand-alone dataset can be misleading.
  - Communities need to mature toward metrics that identify and incentivize ways to improve disciplinary FAIRness.
  - "An appropriate level of FAIRness" is undefined. If FAIR data is a goal for a given research community, community-wide goals should be agreed upon and set.
- Disciplinary Specificity:
  - Communities need to test emerging FAIR data stewardship, FAIR metrics, and data management planning tools to ensure they produce results that are

[32] https://www.go-fair.org/fair-principles/fairification-process/

meaningful in disciplinary contexts.

- Each discipline (and potentially each project) necessarily has different goals and yet commonality and sharing is eventually desired. Standards addressing the broader need for sharing and interoperability should be developed and communicated more broadly.
    - Something like a "Scientific Variable Ontology" may help with global communication among disciplines
    - Meta-data hierarchies may also be useful to help with interoperability and interdisciplinary interpretation

- Training:
    - Participants identified needs for more data curator, researcher and developer training, particularly in the US. Training strategies that scale will be integral to achieving the aspirations of FAIR.
    - Training or expanding knowledge in disciplinary ontology modeling will be important. Very few people have appropriate expertise in this area, yet it is crucial to interoperability.
    - To effect cultural change, FAIR training should be provided for funders' reviewers and awarded research project evaluators, as well as for publication reviewers/editors.
    - Training materials on FAIR principles themselves are not FAIR. FAIRifying Research Data Management training repositories of reusable training materials, such as those from the Data Management Training Clearinghouse, will pay dividends both as exemplars of FAIR assets but also in encouraging self-training and re-mixing of training materials for particular audiences and to meet disciplinary data stewardship training challenges.
- Communities: Although the discussion and activity around FAIR is growing, cohesive communities have not yet formed in the US around grassroots momentum. This is an apparent contrast with the enthusiasm elsewhere, especially in Europe. This lack of peer support may slow efforts at widespread FAIR adoption.
- Conflicting Mandates: Funders' and publishers' Data Sharing and Open Data mandates are already driving data sharing. At the same time, disciplinary uptakes of aspects of the FAIR principles are driving changes in repository and publication system features in ways meant to improve machine actionability on data and metadata. System priorities and different levels of maturity can make for idiosyncratic and frustrating data sharing experiences for the researcher. This is necessarily a symptom of the early stages of FAIR adoption, but that doesn't make it less painful.
- Funding Models: researchers, funders and administrators of all stripes want to know the cost benefit of FAIRifying data and systems. Right now, this is a very difficult thing to provide.

Aftermath & FAIR Futures

We've heard the term creolization[33] from GOFAIR and others. As a FAIR culture emerges what will make it sustainable so FAIR advantages have impact, and what will FAIR implementations look like? Here, for completeness and for a look at future trends, we list some of the activities and workshops that have taken place in the months following our workshop. These represent some perspectives on the current state of FAIR adoption, enabling the process, and prospects for the future.

Data Steward Perspective: US and Abroad

From the perspective of an early career data steward in the United States, Mikala Narlock[34], Digital Collections Librarian, University of Notre Dame, shares that:

> *Due to the diverse mix of representatives present, the MPS FAIR Hackathon provided several unique opportunities for Data Stewards. In particular, this included the chance to discuss data steward needs with these different user groups. For example, this included clear articulations and presentations of materials and physical scientists' needs and victories thus far. One key takeaway was the need to continue supporting and developing FAIR metadata, FAIR metrics, and the capability to appropriately assess FAIR outputs. Additionally, the breakout sessions were valuable opportunities to collaborate with fellow Data Stewards to better identify our own gaps. A key theme that emerged was the need for advanced training opportunities, and the potential for this training to extend to early-career scientists to better educate and inform.*

From the perspective of an Dutch data steward in the EU, Maria Cruz[35], reflected on the hackathon experience in a blog post[36] and at a presentation to colleagues[37] upon her return from the workshop. In summary, Cruz identifies :

> . . . there is still too much of a gap between the ambitions of FAIR - as presented in the workshop, which I found very inspiring - and what I hear and see day to day as I speak with researchers about their RDM practices. Most never heard of FAIR, licenses, persistent identifiers, etc. Most don't share their data or share via personal contacts via email, dropbox, etc. I guess it's a reminder to myself and others that more needs to be done to bridge this gap. And the MPS FAIR Hackathon was one of those initiatives. That's also why I joined the Library Carpentry sprint to produce TOP 10 FAIR Things for Astronomy. And I intend to do more at my university to raise awareness of the FAIR principles.

---

[33] M. Thompson, K. Burger, R. Kaliyaperumal, M. Roos & L.O. Bonino da Silva Santos. Making FAIR easy with FAIR tools: From creolization to convergence. Data Intelligence 2(2020), 87–95. doi: 10.1162/dint_a_00031

[34] Mikala Narlock, Digital Collections Librarian, University of Notre Dame https://directory.library.nd.edu/directory/employees/mnarlock

[35] Maria Cruz, Community Manager Research Data, Library of the VU Amsterdam, https://ub.vu.nl/en/education-research/research-data-services/index.aspx

[36] M. Cruz, "There are great ambitions behind FAIR data but researchers are not on board with it yet," *Open Working: An Experiment in Open Working from 4TU.Centre for Research Data & TU Delft Research Data Services*, 29-May-2019.

[37] M. Cruz, "The NSF MPS FAIR Hackathon--My Views and Reflections." TU Delft Data Champions meeting, 21 May 2019. https://doi.org/10.5281/zenodo.3233530

Chemistry data sharing for publication perspective

Subsequent to the MPS workshop, MPS hackathon advisors, Leah McEwen & Vincent Scalfani co-organized the March 29-30, 2019 FAIR Publishing Guidelines for Spectral Data and Chemical Structures workshop supported by funding from NSF[38]. Later on April 4, 2019 MPS hackathon participant Ian Bruno presented on the results to the Research Data Alliance's Preservation Tools Techniques and Policy Interest Group session at the RDA 13th Plenary Meeting[39] .

Library Carpentry sprint to produce TOP 10 FAIR Things for Astronomy

MPS hackathon participant Maria Cruz and co-convener Meyers met at the MPS workshop and together jointly participated in a Mozilla Spring 2019 sprint to produce the TOP 10 FAIR Things for Astronomy published September 6, 2019.[40] The document is structured into informative small pieces of text (so-called "Things") to jump-start activities a researcher can do to make their data and software more FAIR. They do not have to be followed in a particular order; learners can just pick and choose. The "Things" are sorted under the respective FAIR category to which they belong. This work has been presented at The CODATA-Helsinki Workshop on FAIR RDM in Institutions at the National Archives of Finland on 20-21 October 2019[41] and at the Libraries for Research Data Interest Group session at the Research Data Alliance's 14th Plenary Meeting in Espoo, Finland.

The FAIR Funder pilot programme

The FAIR Funder pilot programme[42,43] aims to make it easy for funders to require and for grantees to produce FAIR Data. Participating funding organizations are together exploring new technologies to define at the time that a request for proposals is issued the minimal set of machine-actionable metadata that they would like investigators to use to annotate their datasets, to enable investigators to create such metadata to help make their data FAIR, and to develop data-stewardship plans that ensure that experimental data will be managed appropriately abiding by the FAIR principles.

Leveraging PIDs and MaDMPs for Interoperability in Research Information Systems

Since the MPS hackathon, inspired by the National Science Foundation's May 20, 2019 Dear Colleague Letter: Effective Practices for Data, the Association of Research Libraries, the California Digital Library (CDL), the Association of American Universities (AAU), and the Association of Public and Land-Grant Universities (APLU) convened a Dec 11-12 2019 conference on "Implementing Effective Data Practices: A Conference on Collaborative Research Support[44]" The event brought together stakeholders to advance understanding among universities and funders about actualizing the FAIR principles in university research information systems, particularly as they relate to permanent unique identifiers (PIDS) and Machine Actionable Data Management Plans (MaDMPs). MPS hackathon

---

[38] https://iupac.org/event/fair-publishing-guidelines-for-spectral-data-and-chemical-structures/
[39] Bruno, Ian. (2019, April). FAIR Chemical Data. Zenodo. http://doi.org/10.5281/zenodo.2642799
[40] https://librarycarpentry.org/Top-10-FAIR//2019/09/06/astronomy/
[41] https://conference.codata.org/Helsinki-CODATA-2019/sessions/167/paper/609/
[42] https://arxiv.org/abs/1902.11162v2
[43] https://osf.io/b9fz4/
[44] https://bit.ly/2MUlqOg

co-convener Meyers was a co-organizer of the later Dec 2019 event. Lessons learned from the prior February MPS FAIR hackathon streamlined and focused the subsequent event's organization and planning. Participants at the Dec 2019 event were aware of how FAIR data culture can propel scientific awareness, discovery and invention. Researchers, data stewards and funders in attendance, asked each other some of the same questions raised at the MPS hackathon months earlier: "What are the next steps to start FAIRification?" "How can universities and funders prioritize what's next in a way that directly benefits research activity and advances research?" "Where can data stewards get needed training?" "How can universities embed FAIRness training opportunities for students and Early Career Researchers?"

FAIR/DataStewardship Training in the United States

Where data stewards can get training and how FAIRness training can be made available at universities going forward are questions that need to be answered in ways that can scale to meet demand and funder expectations for FAIR outputs. Subsequent to the MPS FAIR workshop, and of potential interest to physical sciences stakeholders, there have been at least three related opportunities to further develop FAIR understanding and skills:

- Drexel-CODATA FAIR-RRDM Conference[45]: 31 March-1 April 2019, Drexel University, Philadelphia, PA. At which a summary of this MPS FAIR hackathon was presented to an international audience[46].
- FAIR Data Stewardship Training: 28-31 May 2019, San Diego Supercomputer Center, La Jolla, CA. At which MPS FAIR hackathon convener Meyers and expert advisor Hoebelheinrich attended.
- Advancing FAIR and GO FAIR in the U.S.[47]: 25-27 February, 2020, Georgia Tech, Atlanta, GA. At which MPS FAIR hackathon convener Meyers and participant Narlock will attend. This event will provide advanced GO FAIR training aimed at developing a pool of trainers based in the U.S.. A major goal of this training is to facilitate the development of a community of practice for FAIR awareness and capacity-building in the U.S.. The workshop aims to provide twnety-four partiicpants with preparation for teaching or supporting FAIR data management at their home institution, agency, or professional organization. This model may be scalable if enough subsequent trainings can be offered to create a critical mass of trainers and capable data stewards to support expectations for ubiquitous FAIR data sharing from federally funded research projects. Topics will Include:
    - Adding value to data with semantic tools, and publishing FAIR data points
    - How to teach FAIR techniques, and discussion of the most difficult aspects of teaching FAIR
    - Broader aspects of the Data Stewardship landscape and assessing the uptake of FAIR awareness and practices

Conclusions

The MPS Fair Hackathon provided one of the first forums in the US where researchers from the physical sciences could gather together with experts in FAIR principles and FAIR evaluation tools to participate in hands-on activities related to how to make data FAIR. Researchers brought their own

[45] https://conference.codata.org/Drexel_CODATA_2019/programme/

[46] Hildreth & Meyers. (April 1, 2019) **Implementing FAIR practices and metrics in Physical Sciences research communities**. delivered at Drexel-CODATA FAIR-RRDM Conference Philadelphia, PA

[47] https://www.sdsc.edu/services/data_science/research_data_services.html

data and data problems, making this a concrete exercise. For many, this was their first chance to confront the challenges presented by the process of making data and research results conform to FAIR principles. Throughout the workshop, the participants were asked to generalize the problems they faced in FAIR-ifying their data so that a broad set of “gaps” in infrastructure, training, etc. could be derived as a product of the workshop. These are listed in detail in the preceding sections. A clear focus emerged on training, common practices, and the common infrastructure as the primary needs to be met so as to make the application of FAIR principles easier, and, hence, more common in the physical sciences. There is a tremendous amount of activity around FAIR principles at this time, which makes it both exciting and confusing for researchers. Workshops like this one will be important going forward to provide connections between FAIR experts and researchers who wish to share their data “properly” and are eager for finding the tools and training to do so. As one of the participants said, “the research community IS thinking about (and part of) culture change - one hackathon, one meeting, one dataset and one tweet at a time.” Collecting and channeling this enthusiasm will be important in sustaining the momentum behind the adoption of FAIR principles and the development of research communities around their implementation.

Appendix I: Workshop Registrants

Alshaikh	Ali	University of Alabama
Audus	Debra	NIST
Bartolo	Laura	CHiMaD Northwestern University
Berta	Margaret	University of Notre Dame
Boes	Jacob	Stanford University
Bolton	Evan	NCBI/NLM/NIH/HHS
Bowman	Sara	Center for Open Science
Bruno	Ian	Cambridge Crystallographic Data Centre (CCDC)
Campo	Eva	NSF (Observer)
Chalk	Stuart	University of North Florida
Cruz	Maria	Delft University of Technology
Dillman	Allissa	NCBI/NLM/NIH/HHS
Fokianos	Pamfilos	CERN
Habermann	Ted	Metadata Game Changers
Haghighatlari	Mojtaba	University at Buffalo
Hanif	Hammad	George Mason University
Hanisch	Robert	NIST
Hildreth	Mike	University of Notre Dame
Hoebelheinrich	Nancy	Knowledge Motifs/ESIP
Ladino	Cassandra	USGS
Li	Qingliang (Leon)	NIH/NLM/NCBI
McEwen	Leah	Cornell University
Meyers	Natalie	University of Notre Dame
Mons	Albert	GO FAIR
Narlock	Mikala	University of Notre Dame
Patel	Shrayesh	University of Chicago
Pfeiffer	Nici	Center for Open Science
Plale	Beth	NSF (Observer)
Plante	Ray	NIST
Poirier	Lindsay	University of California Davis
Proffitt	Mason	University of Washington
Publico	Perry	Montgomery College

Pullen	Ian	Purple Polar Bear
Robinson	Erin	Earth Science Information Partners (ESIP)
Scalfani	Vincent	University of Alabama
Schultes	Erik	GO FAIR
Stall	Shelley	American Geophysical Union
Strawn	George	NAS
Trzcinska	Anna	CERN
Tsanaktsidis	Ioannis	CERN
Van Arkel	Annik	Purple Polar Bear Tech Agency
Watts	Gordon	University of Washington
Weston	Joseph	Delft University of Technology
Winther	Kirsten	SLAC National Accelerator Laboratory

Appendix II: Participants' Pre-workshop Questionnaire & Responses

Available online at :
Michael Hildreth & Natalie K Meyers et al. 2019. Pre-Hackathon Questionnaire & Responses.
DOI:https://doi.org/10.17605/OSF.IO/E2F3A

Appendix III: Workshop Agenda

Much more information can be found at the workshop web site: https://mpsfair.crc.nd.edu,

with more resources available at **https://osf.io/km8db/**

Day 1 Feb 27th, 2019

8:00	Breakfast/ Registration	
9:00	Session 1	Introduction, View from NSF, Workshop Goals and Procedures - Mike Hildreth | Beth Plale
9:20	Intros	Participant's brief introductions
9:45	Session 2	Status of FAIR in the US & Overview of efforts Worldwide: ● Intro to FAIR WHY in the US and Abroad, Natalie Meyers ● Enabling FAIR Data and Share Data, Cite Data, leads to Credit and Attribution, Shelley Stall ● Status of FAIR in the US and Efforts Worldwide, Erik Schultes
10:00	Discussion	Questions & Discussion with Beth, Mike, Shelley, Natalie & Erik
10:25	Icebreaker	FAIR Metadata game icebreaker - Ted Habermann
10:30	Coffee	*Coffee available served during the ice breaker*
11:10	Demos	● DMT Clearinghouse - N Hoebelheinrich (5 min) ● Aspects of FAIR Crystallographic Data- Ian Bruno (15 min) ● FAIR Evaluator Demos - Erik Schultes: FAIR Maturity Indicators and Purple Polar Bear: FAIR Tool Manager (40 min)
12:30	Lunch	Luncheon served onsite
13:00	Demos, continued	FAIR Evaluator - Ted Habermann
13:20	Session 3	GOFAIR Matrix and Consortium: Erik Schultes and Albert Mons
14:00	Use cases	● Making Data Interoperable: Use Case - Albert Mons ● Making Data Interoperable: PubChem Demo/Use Case - Evan Bolton
14:45	Evaluator	Chem Phys MatSci/Chem Metrics/Evaluator Tools Trainers

	Breakout	
15:45	Break	
16:00	Demo	Making Data Interoperable: GoldBook of IUPAC Demo: terminology -API access Stuart Chalk
16:30	Discussion	Making Data FAIR Discussion and ad hoc demos Share hacks, tips, tools

Day 2 Feb 28th, 2019

8:00	Breakfast	Breakfast on your own
8:30	Keynote	Day Two Keynote "The Internet and FAIR Data" - George Strawn
9:15	Discussion	• Keynote Discussion - Moderator Mike Hildreth • MPS FAIR Hackthon Gap Analysis and Final Report Rationale - Hildreth • Questionnaire Results - What we've brought/got that we can FAIRify - Meyers • Drexel CODATA March 31-April 1 FAIR and Responsible Research Data Management Workshop Meyers & Schultes • Departure Travel Logistics - RideSharing
9:40	Demo	Physics - FAIR Ecosystems Demos (from among INSPIRE High Energy Physics information system, CERN Document Server, Reana Reusable Analysis and CERN Analysis Preservation Portal) - A. Trzcinska, I. Tsanaktsidis, P. Fokianos
10:00	Breakouts	FAIRifying your data or code : Day Two Breakouts begin
11:30	Icebreaker	Regroup for Discussion & Demonstrate outputs
12:00	Lunch	*Grab and go lunches available*
12:30	Discussion	Demonstration and discussion of outputs continues
12:50	Wrap-up	Wrap Up: MPS Hackathon Report next steps Departure Travel Logistics
13:00	Departure	
13:00	Breakout	(to 15:00) ***Breakout time and consults available for those who want to continue working***