Information Services & Use 40 (2020) 71–85
DOI 10.3233/ISU-200083
IOS Press

Implementing FAIR data for people and machines: Impacts and implications - results of a research data community workshop

Joshua Borycz[a,*] and Bonnie Carroll[b]

[a] *Librarian for STEM Research, Vanderbilt University, Nashville, TN, USA*

[b] *BRDI/CODATA, Information International Associates (IIa), P.O. Box 4141, Oak Ridge, TN, USA*

Abstract. The *Implementing FAIR Data for People and Machines: Impacts and Implications* workshop was organized by the Board on Research Data and Information of the National Academies of Sciences, Engineering, and Medicine (NASEM), the CENDI Federal Information Managers Group, the Research Data Alliance (RDA), and the National Federation of Advanced Information Services (NFAIS), and held at NASEM's Keck Center in Washington, DC on September 11, 2019. The goals of the *Implementing FAIR Data* workshop were to discuss the current status of FAIR data implementation, share what is being done to encourage scientists to share data in machine-readable formats, and examine the implications of FAIR data implementation for people and machines. FAIR data policies, tools, and measures of FAIR data compliance were considered from multiple perspectives. Marcia McNutt, President of the National Academy of Sciences (NAS), offered opening remarks, and the keynote address was presented by Barend Mons, Professor of Bioinformatics at Leiden University Medical Center and President of the International Science Council's Committee on Data (CODATA). Three panel discussions addressed (1) the perspectives of scientists and administrators from U.S. federal agencies, (2) case studies on the implementation of FAIR data practices, and (3) principles and methods of measuring FAIR data compliance. The automation of scientific workflows was discussed by Stuart Feldman, Chief Scientist of Schmidt Futures, a philanthropic organization devoted to investing in research, technology, and science. The workshop closed with highlights and takeaways from each session as summarized by the moderators, followed by general questions.

Keywords: FAIR data, machine learning, research data, open science, data management, data sharing, data metrics, data compliance

1. Introduction: Machine readability in scientific research

Scientific progress largely depends on the ability to build on an existing base of trustworthy knowledge. Tracking the authorship and publication of new ideas has been an important part of this progress because it has provided an incentive for innovators to carefully evaluate methods and results before publishing, to be detail-oriented, and to share their work with their peers. Dissemination for the purpose of critical analysis is one of the primary reasons that the scientific practice has led to such rapid progress in modern history. In fact, this practice has been so successful that researchers have been reluctant to adapt it to

[*] Corresponding author: Joshua Borycz. E-mail: joshua.borycz@vanderbilt.edu.

the rapidly changing technological infrastructure. Scientists largely rely on their own intuition and effort, rather than machine learning, to do the detailed work of gathering and analyzing data. Journal articles are still viewed as the most important products of the scientific process and data is rarely shared [7,24]. This remains true despite the fact that the amount of scientific data produced every year has grown far more rapidly than the number of scientists within each field [3]. Inevitably, a large proportion of very expensive, high quality research data are either discarded or remain unused because the individual scientist has lost the ability to read and digest all that is produced in a field [2,51]. To address this growing challenge, Mark Wilkinson, a leading thinker in the data management community, makes the argument for FAIR (Findable, Accessible, Interoperable, and Reusable) data [52]. The FAIR data principles were designed to help fully integrate big data analytics and artificial intelligence tools into the scientific process. There are many challenges that must be addressed to accomplish this end because most data are not formatted or curated to be "understandable" by machines. This translation process will require a great deal of organizational effort and agreement at the human level before machines can begin to operate directly on data without human intervention. The FAIR principles were designed to serve as the foundation for the shift to machine-assisted and machine-driven research [52].

2. Opening remarks: Open science by design

Marcia McNutt has been a passionate advocate for open science for many years. Since her early days as an oceanographer, working at the Scripps Institution of Oceanography, she noticed a distinct culture of ”data-hoarding” in which scientists would hide data until it was advantageous to their careers to release it. This same culture would often try to encourage others to collaborate only to take their research ideas and use them without appropriate recognition. In the past several decades, oceanography has become one of the leading fields in scientific openness and data sharing. McNutt was involved in improving data sharing practices when she had a role as metadata specialist for marine scientists long before the FAIR principles were codified. She has seen the scientific culture change for the better. Her experience and research have taught her that changing the culture requires developing tools and infrastructure that make adopting open science practices easy for researchers. Tools have to be intuitive, automatic, and usable on multiple platforms. McNutt observed that scientists should be rewarded for adopting open practices as much as they are for publishing in highly cited journals. Universities need to begin paying attention to how scientists share their work as much if not more than where they publish it. The entire research process should be built on the FAIR data principles with openness in mind. These concepts of *Open Science by Design* are derived from a recent report from the National Academies of Sciences, Engineering, and Medicine (NASEM), [25] which argues that "openness and sharing of information are fundamental to the progress of science and to the effective functioning of the research enterprise".

3. Keynote and introduction: The internet for social machines

McNutt mentioned that integrating openness into a scientific community can be done by developing tools that make researcher's lives easier, but it is also necessary to educate researchers on the needs that the FAIR principles address. To that end, Barend Mons spearheaded GO FAIR, an organization designed to encourage the quick adoption of the FAIR principles into mainstream science. The three pillars of GO FAIR are (1) GO CHANGE the culture of science to be more open, (2) GO TRAIN data stewards that can help guide scientists, and (3) GO BUILD infrastructure that allows scientists to easily adopt FAIR data practices. It may seem like GO BUILD is the only one of the GO FAIR pillars that relates to machine-readable data, but most data is still generated by humans, and it is humans that must ultimately translate their work into a language that machines can understand [16]. This translation requires turning ambiguous concepts into curated bits of information.

Concepts can be understood by machines by linking bits of information into networks and finding common patterns. Two main questions arise: What information is necessary for a machine to understand what a FAIR data record means? What is the minimum machine-readable record such that a machine could understand the concepts within it and related to it? Mons stated that the fundamental pieces of information a record needs to be FAIR are (1) the type of record (e.g., temperature measurements, survey answers, images), (2) the operations that are possible, and (3) the operations that are allowed. With this information, commonalities between FAIR digital objects (DOs) could be found and concepts could be linked together more easily. If the types of data and operations within FAIR DOs were understood and standardized then linking DOs across networks would be simpler. In other words, if each FAIR DO in a network had sufficient metadata, then machine learning could be used to link and analyze large sets of data for patterns that humans could easily miss.

Mons refers to the metadata related to concepts within fields as *knowlets*. Connecting *knowlets* into networks could allow machines to connect similar concepts in different fields, link concepts that have been written in different languages, and ultimately to infer complex links between concepts over thousands or millions of papers. This is how the internet of social machines could be used to the advantage of science. However, connecting these *knowlets* is incredibly difficult. Apart from the lack of incentives to upload FAIR data, publishers have different rules on the publication of supplemental data, rights to access data are inconsistent, and the information on the rights themselves is often difficult to find [15]. To take advantage of the power of machine learning, a culture of data sharing must be encouraged in the short term. This must be done by making it easy to share data as McNutt suggested, through regulation, and through positive or negative reputation-based incentives. In closing, Mons clearly stated that it is unethical for researchers to be paid by taxpayers and not share their research data. To make full use of the public investment in research, data must be FAIR.

4. Perspectives from U.S. Federal Organizations

Making it easy to share data will certainly improve practices, but only if there are monetary or career incentives created that are proportional to the level of effort required. Since U.S. federal agencies are responsible for funding massive amounts of research, federal initiatives are needed to change the culture. In this session, Michael Huerta, the Associate Director for Program Development in the National Library of Medicine (NLM), an institute of the National Institutes of Health (NIH), Beth Plale, Science Advisor for Public Access at the National Science Foundation (NSF), Robert Hanisch, Director of the Office of Data and Informatics (ODI) at the National Institute of Standards and Technology (NIST), and Laura Biven, Program Manager of Advanced Scientific Computing Research (ASCR) at the Department of Energy (DOE), spoke about their agencies' contributions to open science and FAIR data.

4.1. National Institutes of Health

Huerta opened his remarks by clarifying the distinction between data science and open science. Data science focuses on using tools, whereas open science is about paradigms. Data science tools are typically developed for specific purposes within certain fields. For open science to take hold in each of these fields, scientists must adopt practices that will not necessarily help them achieve their specific, immediate tasks, but that will be helpful in the long term. This has led many agencies to put in place open data initiatives. Unfortunately, these tend to remain disconnected from initivatives in other agencies. A more promising approach is to encourage all agencies and all scientific groups to adhere to the FAIR principles. This is the primary goal of the NIH's and NLM's strategic plans for open science.

NIH is a large research organization with over 5,000,000 users that distributes 115 Tb of health data each year. NLM, the largest biomedical research library in the world, is responsible for housing and distributing these data. The key objective of the *NIH Strategic Plan for Data Science* is to promote FAIR data sharing for all NIH research [28]. This entails providing FAIR-enabled, open access to all datasets supporting publications funded by NIH. NIH encourages all researchers to place data in domain-specific repositories as a first choice. Other options include uploading up to 2 GB of data along with a published paper, uploading up to 20 GB in an NIH repository, or uploading petabytes of data to the cloud using STRIDES (Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability Initiative). STRIDES was developed in partnership with Google Cloud and Amazon Web Services [30]. As a subset of the NIH, the three goals of the NLM Strategic Plan are (1) connect and enhance resources through computational curation and critical valuation of resources, (2) optimize the user experience by enhancing outreach and engagement, and (3) offer data science and open science training to employees, collaborating organizations, and users [38]. The pillars of NIH's strategic plan focus on data stewardship, workforce development, data management tools, modernizing data platforms, and improving infrastructure by storing data in the cloud.

4.2. National Science Foundation

NSF has been changing its approach to scientific support and funding. Beth Plale serves as an advisor at NSF and works to provide public access to NSF-funded research, to encourage good data management practices, and to ensure that research data is shared in machine-readable formats by working closely with GO FAIR. One of NSF's boldest initiatives has been the *NSF Public Access Repository* (NSF-PAR), which makes all published work and data that were developed using NSF funding available for free [40]. NSF-PAR is a repository centralized at NSF that is constructed using the DOE PAGES template [8]. All award recipients must deposit the author manuscript of a publication as part of the reporting process [39]. This has recently been extended to include reports of workshops funded by NSF. In 2019, NSF produced a letter outlining its support for open science principles as well as guidelines on using globally unique persistent identifiers for research data and machine-readable data management plans (DMPs) [41]. NSF is working in coordination with other agencies based on a recent effort to encourage open science for all research funded with public money [25]. The *Open Science by Design* consensus study outlines the current state of open science, the vision for the future of open science, and how to transition from one to the other quickly. NSF is also coordinating with GO FAIR to improve the metadata within its *Big Data Innovation Hubs* [43,44].

4.3. National Institute of Standards and Technology

Robert Hanisch (NIST) has been working within the Office of Data and Informatics (ODI) to bring about cultural changes related to data sharing [33]. Much like NSF and NIH, NIST requires that papers derived from NIST-funded research be accompanied by data that is compliant with FAIR principles [32]. NIST offers four areas of active support to researchers through ODI. These include (1) curating and making available standard reference data (SRD) for exploratory research (49 free SRD databases) [37], (2) providing research data support through training and tool development, (3) providing data science infrastructure that includes cloud services and research domain specific expertise, and (4) engaging with other data related communities and organizations like the Research Data Alliance (RDA), CODATA, and the National Data Service (NDS). Along with these organizations, NIST is supporting the institution of the FAIR principles to maximize the return on federal research investment and address reproducibility issues within science. ODI has been a strong leader in the effort to improve research reproducibility and it continues to be one of the primary drivers of their work [17].

Hanisch spoke about some specific tools that NIST supports, such as the *NIST Materials Data Repository*, which is designed to help communities work across organizational boundaries on data that is not yet ready for publication [34]. The *NIST Data Discovery Platform* is a searchable interface for all data and resources provided through NIST [35]. One goal of ODI is to encourage the use of Laboratory Information Management Systems (LIMS) so that FAIR data and metadata are captured at the moment of observation or, ideally, directly from the instruments themselves. NIST is also developing a Research Data Framework (RDaF) that will be used as an overall guide to the actors and stakeholders involved in creating or using research data [36].

4.4. Department of Energy

DOE is at the forefront of the movement to use machine-learning to increase the rate of advancement of scientific research [46]. DOE has recently built the Summit supercomputer at Oak Ridge National Lab (ORNL), currently the most powerful supercomputer in the world [12,50]. Laura Biven emphasized the important part that DOE plays in encouraging FAIR data practices. DOE is integrated within the scientific community because of its high-performance computing resources and 17 national labs, and it is one of the primary funders of research in the United States, which means that DOE can greatly influence the data sharing practices of scientists. This bodes well for the scientific community because DOE has been a leader in providing public access to research through its DOE PAGES portal, available since 2014 [8,9]. Currently, the DOE Office of Science funds approximately 40% of the scientific research done in the United States [10]. The DOE Office of Science has recently been organizing discussions on how to advance and support machine-learning within science. The Office of Science wants to develop a cartography (i.e., a map of related data) based on relationships among data within their repositories, to see what gaps remain for the implementation of machine-learning processes. Biven stated that DOE is uniquely positioned to lead the use of machine-learning in science as its mission demands paying attention to and adapting new technologies for multiple purposes [11,13].

4.5. General discussion on federal perspectives

The many efforts of NIH, NSF, NIST, and DOE inspired an interesting discussion among the speakers and members of the audience. One of the first issues mentioned by Huerta was that agency efforts often progress on their own without outreach to help align efforts with other institutions. There were some concerns expressed about this issue by the audience, but it was clear that each of the panel members was committed to being open about its practices and to using FAIR data practices, which should make their research data efforts trasparent and usable across institutions. Another important discussion related to updating old observational data within these institutional repositories according to the FAIR principles. The general consensus was that this might be done for some important datasets, but that most funding is directed towards new data rather than old. Plale mentioned that even if older scientific data could not be made FAIR, the metadata records themselves could be greatly improved.

The idea of rewarding scientists for sharing research openly was also raised for discussion among the panelists. Hanisch pointed out that the return on investment is already there in fields such as astronomy, in which archivists look at and annotate highly detailed images of stars and planets to generate research findings. Plale said that NSF funds many small grants and creates community resources that encourage FAIR data practices. One point mentioned by an audience member was that librarians at research institutions have been promoting FAIR data practices and trying to change publishing incentives at universities for many years. It was suggested that some of their efforts and ideas should be taken into account by federal agencies. This connected to the next discussion on how funding is distributed to encourage data stewardship. Each of the panelists stated that their agencies were beginning to hire and distribute funds to information specialists such as librarians. Hanisch stated that NIST already works very closely with librarians, but Huerta, Plale, and Biven felt that the cohort of information specialists at national laboratories is still small and more investment is needed.

5. Case studies: Implementing FAIR networks

Understanding the current use and impact of FAIR data practices on scientific research is vital because it will provide insight on how to proceed within the diverse range of fields and institutions that drive scientific practice. This session, which focused on large FAIR initiatives, revealed how practicing scientists actually interact with their data and how the practices and methods of motivating scientists to engage with the FAIR principles might be changed for the better. Giridhar Manepalli, Director of Information Management Technology at the Corporation for National Research Initiatives (CNRI), Rebecca Koskela, Executive Director of DataONE, and Larry Lannom, Director of Information Services and Vice President at CNRI spoke about their implementations of the FAIR principles.

5.1. An architectural approach to FAIR digital objects

Manepalli's presentation focused on network architecture as an engine for implementing FAIR DOs. Manepalli summarized the end goal of FAIR as allowing computers to produce results from DOs at a rate faster than humans without much human input. This issue has been solved in some areas. For example, mobile weather apps use automated radar to predict weather, send the data to a server, and then automatically send the data to app users' smartphones. There is very little interaction with humans in this process. Areas in which machine-learning is still not a tractable solution include email servers. Emails contain ambiguous headings, expressions, different languages, and inside jokes, which makes them almost impossible to control automatically. How can ambiguous environments such as this be automated? The approach at CNRI is to switch from a system-centric platform to an information-centric network based on DOs. This would allow humans to interact with DOs as needed while still having most processes automated.

The three pillars of the information-centric approach are (1) assigning unique resolvable identifiers to DOs that are free of semantic ambiguities and are long-lasting, (2) using DOs with identifiers that allow easy consumer interaction and interpretation, and (3) providing an interface protocol that helps DO operations remain consistent even upon manipulation by consumers. Manepalli used the analogy of the success of the internet protocol suite for distributing DOs across computers, which has allowed for both automation and human interaction to coexist on the same network successfully. This approach would allow for machine-readable DOs to be distributed as TCP/IP packets across networks and interpreted in personalized human readable formats at various gateways, minimizing the need for complex metadata within the packets themselves.

5.2. *Data sharing perceptions of scientists*

The Data Observation Network for Earth (DataONE) began in 2009 at the University of New Mexico, and was one of the first *data nets*, federated networks of repositories, funded by NSF [23]. DataONE focused on earth, ecology, life, and environmental data. Koskela was the Executive Director of DataONE from its inception and encouraged research that sheds light on the attitudes and perceptions of scientists towards data sharing. There have been three large surveys of scientists carried out by researchers at DataONE [47–49]. The results of these surveys indicate that most scientists (~75%) are satisfied with their short-term data storage practices, while only ~50% are satisfied with the long-term practices in data storage. However, the majority of researchers (68%) do not follow any established community practice when storing their data. Most scientists neither receive sufficient training in data management, metadata assistance, or citation practices, nor seek out the help of librarians or data management experts to help with their data. Researchers tend to contact only their immediate colleagues on data matters. Scientists do not tend to be concerned about sharing data unless they can be guaranteed credit for their work, and many do not feel that they can trust data gathered by others to conduct their own research. On a positive note, funding for data storage has improved and more scientists have become more aware of locations where they can store their research data.

Technical barriers and issues of trust have led Koskela to endorse the *Core Trust Seal* repository certification, which seeks to improve trust in shared research data by guaranteeing a high standard of data quality and openness [6]. Koskela emphasized that joining the *Core Trust Seal* cohort of certified repositories would build stakeholder confidence, raise awareness about digital preservation, ensure transparency, and provide recognition of trustworthiness for all those involved. Koskela also serves as the Executive Director of EarthCube, an NSF-funded geoscience repository, which was the first repository to be certified by *Core Trust Seal* [14]. Koskela's goals for the future are to gather more information on perceptions of data sharing, support training initiatives for research data management, and provide opportunities for quality data repositories to become certified.

5.3. *Research infrastructure for natural science collections*

The Distributed System of Scientific Collections (DiSSCo) repository is a collection of information about existing biological and geological specimens that are widely distributed across museums, universities, botanical gardens, and other institutions throughout Europe [22]. This repository unifies scientific assets under common curation and makes the data more FAIR. Currently, the repository contains data on more than 1.5 billion specimens from 119 collaborating institutions, 5,000 scientists, and 21 countries. DiSSCo was funded in 2019 for an additional three years and 31 new collaborators. Lannom works with DiSSCo to make scientific data in Europe align with the FAIR principles and views the move to FAIR data practices as being vital to the future of science. Currently, there are too many societal problems that need scientific solutions and too much quality data going to waste due to lack of access. Lannom noted that, although FAIR is widely accepted in Europe, there are still many practical issues remaining. Researchers need to have more tools available to easily make their data FAIR. Making data FAIR will broaden research possibilities for future generations. Problems that now seem intractable could easily be addressed using machine-learning later if data openness is given a higher priority now.

5.4. General discussion on case studies

An audience member led with a provocative assertion by stating that attempts to automate scientific processes with machine learning have not had much success in the past and that more success has been gained from making the jobs of researchers easier (e.g., by developing simple data management tools that enable point-and-click metadata creation and data organization) and by creating federal initiatives that require data management and sharing to obtain research funding. Given this history, the questioner asked if technology has changed to the point where automation is more feasible or whether scientists should expect the same difficulties that have plagued previous machine-learning research approaches. Manepalli responded that it is difficult to tell. Human beings have certain flaws with respect to data management and consistency that computers do not, so we should be prepared to switch to automation when it is possible to do so. The way that humans can continue to contribute in meaningful ways is by adding context to data that is compatible with multiple environments. In the short term, this will be more important than constructing DOs that are usable by machines. Lannom stated that advances in computing have been dramatic and that although we may not be ready now, the switch to automated research will come quickly and we should be prepared.

6. Defining and measuring FAIR

The use of metrics within scientific communities to understand how well FAIR is being implemented is important, but has proven to be controversial in some aspects. Metrics like the impact factor have grown to the point where they are overused and provide perverse incentives to researchers to publish flashy results rather than quality, reproducible research. To avoid similar drawbacks, measuring levels of adherence to the FAIR principles should proceed carefully and transparently. Luiz Bonino, International Technology Coordinator for GO FAIR, Maryanne Martone, Professor Emerita in Neuroscience at the University of California, San Diego, and Keith Russell, Engagements Manager for the Australian Research Data Commons (ARDC) spoke on their uses of metrics.

6.1. Core criteria of FAIR principles

Both Bonino and Russell emphasized similar concepts in their discussion on how to apply the FAIR principles to research data in a practical way. Both focused on how the FAIR principles need to be flexible so that different communities can use them in ways that suit their specific needs. Some key components, such as improving metadata and creating unique identifiers for research data, should be applicable in all fields, but each scientific community must be able to decide for itself which standards they require for their published data and which standards work best for their research. Russell emphasized that the FAIR principles cannot be too strict and that assessment should reflect the various needs of different scientific fields. Compelling all scientists to use a single set of FAIR metrics would be a disservice to scientists that work in fields where some data security or restrictions are needed, or where data is much more difficult to organize. Russell proposes that the core stakeholders within many research communities be brought together to discuss assessment methods for the FAIR principles. This will help determine which communities have the infrastructure to support FAIRness and which principles are easiest to implement. For example, interoperability is one of the more difficult principles to implement, as it requires designing metadata and code that can work on many different platforms. In some fields there are too many platforms in use for this to be possible in the near future. Both Bonino and Russell emphasized that the key to addressing these issues is to develop metrics that do not harshly compare researchers or research communities to each other, and that are based on input from members of each research community.

6.2. The state of FAIR in neuroscience

Martone is a prominent member of the neuroscience data community, which has many unique data needs that cannot be easily captured by a broad-stroke FAIR data initiative. Neuroscientific research deals with diverse datasets that include highly detailed, annotated images from instruments, as well as Brain Atlases, which are complex 3D models of brains. At this stage, none of these Brain Atlases have been designed with the FAIR principles in mind [20]. Martone stated that the FAIR principles make sense as concepts that could apply to a wide range of scientific pursuits because they are ideas to be strived for rather than items that can be achieved in full. The Brain Research through Advancing Innovative Neurotechnologies (BRAIN) initiative is funded by NIH and is meant to increase the rate of understanding of the human brain through technological innovation [29]. The BRAIN initiative seeks to incorporate the FAIR principles into its research methodology, but the standards for doing so are not yet well defined [31]. Martone mentioned the Neuroinformatics Coordinating Facility (INCF), which is a community platform and standards organization based in Stockholm, has been supporting the development of neuroscientific standards with some success. The successes of INCF stem from the guidance/training materials and tools that it provides on the platform, a set of consistent criteria that support FAIR principles, a focus on incorporating community input, and its ability to interface with the broader health community [19].

To improve the extent to which neuroscience embraces the FAIR principles, Martone suggested that data experts spend some time training people that are not directly involved in research but still produce important data (e.g., technicians and nurses) and find funding mechanisms to support this training. Martone also mentioned that neuroscientists often make use of distributed data repositories like *dkNET*, a search portal funded by NIH's National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), which should allow for the easy implementation of FAIR practices and advertisement of best data practices and training opportunities [27]. In closing, Martone stated that assessment and metrics can backfire if put into force before they are vetted. For now, FAIR-minded people should focus on understanding how the FAIR principles can be applied within the resource constraints of the various fields.

6.3. General discussion on FAIR metrics

The audience and speakers discussed which organizations or communities should be responsible for defining good data practices, issues with interoperability in science, and the dangers of using metrics.

In Martone's view, decisions regarding FAIR data practices need to be made at multiple levels. Broad suggestions can be made by institutions, but these decisions cannot be so inflexible as to limit domain-specific needs. Each scientific community must be in charge of the day-to-day data standards that it implements. This is why incentives and easy-to-use tools are so important to FAIR practices. Martone suggested targeting neuroscience communities that already have incentives to share data and developing infrastructure for their specific needs. The need to put communities in charge of their own standards is what makes interoperability such a difficult FAIR principle to implement. Bonino suggested that the solution to this issue may be to focus on making the metadata interoperable rather than the data itself. Data formats and types vary too widely among fields to be completely defined at this point. Issues often arise when trying to find semantic relationships between different fields because they use different terms and emphasize different meanings for similar terms. Eliminating the use of implicit semantics in metadata is one important way to improve interoperability.

7. FAIR in discovery, search, and scholarly communication

This session focused on how the FAIR principles are actually being used by scientists and how compliance can be monitored. The metrics session outlined some issues with using metrics too broadly and each of the panelists agreed that metrics should not be released for the purposes of comparing scientists to each other. At this point, metrics are simply being used to see which FAIR principles are being applied by which communities and why. Matt Jones, Director of DataONE, shared how the FAIR principles are used by scientists in various fields that store data with the DataONE repository network. Daniella Lowenberg, Data Publishing and Data Metrics Product Manager, University of California Curation Center (UC3), discussed how metrics can be used to provide credit for sharing data properly.

7.1. Quantifying FAIR in the DataONE repository network

The 42 repositories and over 150,000 researchers that use the DataONE repository network share a common data interface. This makes tracking users' data practices quite simple. Jones and the team at DataONE broke the FAIR principles down into a series of simple checks to see which scientists using the repository network were complying with FAIR principles [21]. Some of the checks for FAIR compliance are provided in Table 1. MetaDIG is a metadata-checking tool created by the National Center for Ecological Analysis and Synthesis (NCEAS) that is available on Github. It provides an outline that shows which metadata concepts should be included in the analysis of the data in the repository network [26]. The results of the quantification of compliance with FAIR principles should be seen as comparative rather than punitive. Each scientific community must decide on its own requirements. Furthermore, FAIR compliance should not be seen as binary. Compliance exists on a continuum that is heavily dependent on the scientific field in question.

The checks in Table 1 can be used to determine whether individual researchers or institutions comply with the FAIR principles for datasets on different projects or to distinguish between the needs of different fields. The results so far indicate that the *Reusable* checks tend to have the lowest rates of compliance. Individuals can change their levels of compliance with each new dataset they submit from month to month. The FAIR scores varied widely for each of the repositories within the DataONE network. Each of the FAIR principles were applied with differing levels of compliance to each field and each field presented unique challenges for FAIR compliance. One of the reasons that the FAIR principles were written broadly and

Table 1
Checks for FAIR compliance used on data within the DataONE repository network [21]

Findable	Accessible	Interoperable	Reusable
Title	Distributor	Data format	Metadata license
Metadata identifier	Publisher	Metadata schema	Resource description
Resource identifier	Landing page	Checksum	Semantic links
		Attribute name unique	

somewhat flexibly was so they could be adapted to the needs of any scientific community. These results indicate that community input is needed to determine the necessary levels of FAIR compliance and to interpret the FAIR principles for each unique case.

7.2. *FAIR metric for FAIR data: Making data count*

Lowenberg focused on how credit might be given to those who comply with the FAIR data principles. Currently, sharing data is not really rewarded in academia. This led Lowenberg and her colleagues at UC3 to develop six steps that should lead to reasonable data metrics.

(1) **Value data rather than only value publications.**
(2) **Ensure that transparent infrastructure for data sharing is available to researchers so that metrics are understandable and citations are properly attributed.** This is an issue because most metadata are still not machine-readable, which means that metadata are difficult to find and data citations are rare. This could be solved by integrating Crossref/DataCite data into repositories so that links between funders, publications, and researchers are preserved [18].
(3) **Establish bibliometric principles for data so that meaning can be assigned to data citation metrics.**
(4) **Implement a curation and peer review process for data to ensure quality and trust.**
(5) **Continuously gather information to establish community agreement and provide researchers within each field data management and sharing guidance.**
(6) **Provide data management support to these communities.** Compliance and quality assurance cost money and institutions need to support this infrastructure if they want to see data management improve. Lowenberg stated that some communities have taken a few of these steps, but that funding and motivation are still generally lacking.

7.3. *General discussion on quantification of the FAIR principles*

There were some concerns among audience members that metrics of any kind might lead certain researchers to try to take advantage of the system. For example, some researchers might split a dataset produced from their work into smaller pieces or simply recombine and publish other datasets to inflate their citation rates without doing a significant amount of additional work. However, this is often done with research articles and is generally not done purely to increase citations. There are often good reasons to split up both articles and datasets, especially if doing so makes the publication easier to follow or eliminates data that researchers citing the work may find superfluous. Furthermore, combining, cleaning, and publishing data combined from multiple sources is a good practice that can make that data useful to a wider range of researchers. Since it is difficult to distinguish between legitimate and fraudulent publication behavior, the researchers at DataONE felt that the benefits of a user-controlled, open archive model outweighed the potential negative incentives. For that reason, DataONE decided to use a profile-centered design, allowing researchers to decide what goes in each dataset and associating each dataset with the user who published it.

8. FAIR data and scientific workflows

Thinking of ways to integrate the FAIR principles into the daily lives of scientists is necessary to change the scientific culture. Stuart Feldman from Schmidt Futures has been trying to do this for many years. Feldman began by stating that workflows will be even more necessary for research when machine-learning becomes integrated into the process. For artificial intelligence to make the most of research data, the workflows devised/used by scientists must be documented along with their research data. This way, machine-learning could be used to produce new data very rapidly and to create metadata for new types of research. The combination of FAIR-compliant, automated workflows can produce an abundance of rich, standardized metadata for use in future research. Currently, scientific workflows are not at all standardized within or between fields. There are thousands of workflow engines and languages. Feldman has discovered that this is partially the result of certain qualities in scientists that resist conformity and control.

Feldman also indicated that workflow compliance is changing in someresearch sectors. It is now obligatory to record workflows at some U.S.federal agencies such as the Food and Drug Administration (FDA). Thereare some open workflow systems available that are fairly simple toadopt and have growing user bases [1,4]. Jupyter notebooks have made iteasier to write workflows for coding projects and have been adopted bymany scientific communities [42]. There has also been a generationalshift in coding ability as young researchers in multiple fields havebegun using Python and R programming languages [5]. The technologicalprowess of early career scientists could help drive machine-learning asa research tool.

Feldman talked about three big directions for machine-learning in thefuture. (1) Machine-learning as a tool for research is alreadyhappening. Laboratory automation, natural language analysis, and image-analysis all involve machine-learning. (2) Machine-learning as a driverof research occurs when research projects are executed by machinesafter some initial human input and design. The latter requiresorganizing datasets so that they can be analyzed by machines. Humanintuition will not be as useful for projects driven by machine-learning. Rather, objective, validated measures need to be decided upon before analysis begins so that results are displayed in an interpretable manner. (3) Machine-learning as a creator of knowledge refers to a time when machines are so well trained that they can conceive of, develop, carry out, and analyze research results on their own. This use of machine-learning seems very far off, but Feldman believes that it is closer than we think. Schmidt Futures is working to incorporate machine-learning into astronomy and chemistry research and has funded a NASEM study on the use of advanced and automated workflows in scientific research that is currently underway.

9. Filling the gaps: Conference summary

In the final session of the *Implementing FAIR Data* workshop, the moderators summarized the main points of each session. In general, scientists from U.S. federal agencies, academic institutions, and industry agree that FAIR practices are important and institutions are beginning to find ways to coordinate efforts. There needs to be more emphasis on community input and infrastructure with respect to the implementation of the FAIR principles. Currently, many scientists do not have sufficient support or incentives to adopt FAIR practices and there are reasonable fears about using metrics to measure compliance. Many workshop participants feared that metrics will be seen as punitive by researchers. Furthermore, the infrastructure is not at the point where FAIRness is easy to implement, which is one important reason why many scientists may choose not adopt the principles. In other words, simple tools and clear definitions of what constitute FAIR practices need to be decided upon and outlined for each scientific community. Otherwise, FAIR data compliance will be seen as just another barrier by scientists rather than an active good. In the end, FAIR data does not guarantee open science, but it does help enable it.

The final point in the summary session addressed how to educate researchers on the FAIR principles, ethics in research, and the conflict between the flexibility and sustainability of FAIR. Machines are capable of doing certain types of research rapidly, but they are not capable of incorporating ethics into research. As such, the human element will be necessary in research for the foreseeable future. One audience member noted that bad data can be perfectly FAIR. This means that considerations of research malpractice and prejudice must be accounted for in some type of peer review process. Regarding education, many audience members pointed out that librarians have been working to improve data management practices of researchers for some time and, as such could help researchers understand and implement the FAIR principles. The concern about the sustainability of FAIR if the principles are not clearly defined was addressed by admitting that certain concerns like ethics and education will never be automatable. These principles should remain flexible enough to allow communities to decide which aspects of FAIR should be completed by humans, which by machines, and which aspects FAIR should not address altogether.

Acknowledgements

This authors would like to thank Ester Sztein (NASEM) for carefully editing this article. Thanks also to workshop sponsors for supporting the travel of the first author and the publication of this article, and to Information International Associates for handling the logistical details. Partial funding for this workshop was provided by NIH and NSF awards to BRDI (NASEM), and supported by the CENDI Federal Information Managers Group, RDA, and the National Federation of Advanced Information Services (which merged with the National Information Standards Organization). All workshop presentations can be found on the NASEM website (https://www.nationalacademies.org/event/09-11-2019/implementing-fair-data-for-people-and-machines-impacts-and-implications).

Competing interests

The authors have no competing interests to declare.

References

[1] A.El. Assas, Wexflow - Open source workflow engine. Retrieved December 1, 2019, from https://wexflow.github.io/, 2019.

[2] M. Baker, 1,500 scientists lift the lid on reproducibility, *Nature* **533**(7604) (2016), 452–454. doi:10.1038/533452a.

[3] L. Bornmann and R. Mutz, Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references, *Journal of the Association for Information Science and Technology* **66**(11) (2015), 2215–2222. doi:10.1002/asi.23329.

[4] S. Bowers, T. McPhillips, S. Riddle, M.K. Anand and B. Ludäscher, Kepler/pPOD: Scientific Workflow and Provenance Support for Assembling the Tree of Life, 2008, doi:10.1007/978-3-540-89965-5_9.

[5] S. Cass, The Top Programming Languages 2019 - IEEE Spectrum. Retrieved December 1, 2019, from https://spectrum.ieee.org/computing/software/the-top-programming-languages-2019, 2019.

[6] CoreTrustSeal. *Core Trustworthy Data Repositories Extended Guidance*. Retrieved from https://www.coretrustseal.org/wp-content/uploads/2017/01/20180629-CTS-Extended-Guidance-v1.1.pdf, 2016.

[7] J.L. Couture, R.E. Blake, G. McDonald and C.L. Ward, A funder-imposed data publication requirement seldom inspired data sharing, *PLOS ONE* **13**(7) (2018), e0199789. doi:10.1371/journal.pone.0199789.

[8] DOE. DOE PAGES. Retrieved November 26, 2019, from https://www.osti.gov/pages/, 2014.

[9] DOE. *Public Access Plan*. Retrieved from http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf, 2014.

[10] DOE. About the Office of Science. Retrieved November 26, 2019, from https://www.energy.gov/science/about-office-science, 2019.

[11] DOE. About Us | Department of Energy. Retrieved November 27, 2019, from https://www.energy.gov/about-us, 2019.

[12] DOE. Advanced Scientific Computing Research | Department of Energy. Retrieved November 26, 2019, from https://www.energy.gov/science/ascr/advanced-scientific-computing-research, 2019.

[13] DOE. Office of Science National Laboratories | Department of Energy. Retrieved November 27, 2019, from https://www.energy.gov/science/science-innovation/office-science-national-laboratories, 2019.

[14] EarthCube. *EarthCube Charter*. Retrieved from https://www.earthcube.org/document/2019/earthcube-charter-3-19, 2019.

[15] A. Extance, How AI technology can tame the scientific literature, *Nature* **561**(7722) (2018), 273–274. doi:10.1038/d41586-018-06617-5.

[16] GO FAIR. GO FAIR. Retrieved November 26, 2019, from https://www.go-fair.org/, 2019.

[17] R.J. Hanisch, I.S. Gilmore and A.L. Plant, Improving reproducibility in research: The role of measurement science, *J Res Natl Inst Stan* **124** (2018). doi:10.6028/jres.124.024.

[18] G. Hendricks, R. McFall, E. Pentz, D. Tkaczyk and A. Tolwinska, *Crossref Fact File: 2018–2019 Annual Report* 10.13003/y8ygwm5, 2019.

[19] INCF. Standards and Best Practices organisation for open and FAIR neuroscience | INCF - International Neuroinformatics Coordinating Facility. Retrieved November 27, 2019, from https://www.incf.org/, 2019.

[20] K.A. Johnson and J.A. Becker, The Whole Brain Atlas. Retrieved November 27, 2019, from Harvard Medical School website, http://www.med.harvard.edu/AANLIB/home.html, 1995.

[21] M.B. Jones, P. Slaughter and T. Habermann, *Quantifying FAIR: automated metadata improvement and guidance in the DataONE repository network,* doi:10.5281/ZENODO.3408466, 2019.

[22] L. Lannom, D. Koureas and A.R. Hardisty, FAIR data and services in biodiversity science and geoscience, *Data Intelligence* (2019), 122–130. doi:10.1162/dint_a_00034.

[23] W. Michener, D. Vieglais, T. Vision, J. Kunze, P. Cruse and G. Janée, DataONE: Data observation network for earth — preserving data and enabling innovation in the biological and environmental sciences, *D-Lib Magazine* **17**(1/2) (2011). doi:10.1045/january2011-michener.

[24] T. Miyakawa, No raw data, no science: Another possible source of the reproducibility crisis, *Molecular Brain* **13**(1) (2020), 24. doi:10.1186/s13041-020-0552-2.

[25] NASEM. *Open Science by Design,* doi:10.17226/25116, 2018.

[26] NCEAS. metaDIG., 2019.

[27] NIDDK. dkNET: Connecting research resources. Retrieved November 27, 2019, from https://dknet.org/, 2019.

[28] NIH. NIH Strategic Plan for Data Science | Data Science at NIH. Retrieved November 26, 2019, from https://datascience.nih.gov/strategicplan, 2018.

[29] NIH. Brain Initiative. Retrieved November 27, 2019, from https://braininitiative.nih.gov/, 2019.

[30] NIH. STRIDES Initiative | Data Science at NIH. Retrieved November 26, 2019, from https://datascience.nih.gov/strides, 2019.

[31] NIH. The Neuroimaging Data Model: FAIR descriptors of Brain Initiative Imaging Experiments | Brain Initiative. Retrieved November 27, 2019, from https://braininitiative.nih.gov/funded-awards/neuroimaging-data-model-fair-descriptors-brain-initiative-imaging-experiments.

[32] NIST. *NIST Plan for Providing Public Access to the Results of Federally Funded Research*. Retrieved from http://www.whitehouse.gov/omb/circulars_a110/, 2015.

[33] NIST. Office of Data and Informatics | NIST. Retrieved November 26, 2019, from https://www.nist.gov/mml/odi, 2018.

[34] NIST. Materials Data Repository. Retrieved November 26, 2019, from https://materialsdata.nist.gov/, 2019.

[35] NIST. NIST Science Data Portal. Retrieved November 26, 2019, from https://data.nist.gov/sdp/#/, 2019.

[36] NIST. Research Data Framework (RDAF) Workshop | NIST. Retrieved November 26, 2019, from https://www.nist.gov/news-events/events/2019/12/research-data-framework-rdaf-workshop, 2019.

[37] NIST. (2019d) . Standard Reference Data | NIST. Retrieved November 26, 2019, from https://www.nist.gov/srd, 2019.

[38] NLM. *A Platform for Biomedical Discovery and Data-Powered Health*. Retrieved from https://www.nlm.nih.gov/pubs/plan/lrp17/NLM_StrategicReport2017_2027.html, 2017.

[39] NSF. *Public Access Plan: Today's Data, Tomorrow's Discoveries: Increasing Access to the Results of Research Funded by the National Science Foundation*. Retrieved from https://www.nsf.gov/pubs/2015/nsf15052/nsf15052.pdf, 2015.

[40] NSF. NSF-PAR. Retrieved November 26, 2019, from https://par.nsf.gov/, 2016.

[41] NSF. Dear Colleague Letter: Effective Practices for Data (nsf19069) | NSF - National Science Foundation. Retrieved November 26, 2019, from https://www.nsf.gov/pubs/2019/nsf19069/nsf19069.jsp, 2019.

[42] J.M. Perkel, Why Jupyter is data scientists' computational notebook of choice, *Nature* **563**(7729) (2018), 145–146. doi:10.1038/d41586-018-07196-1.

[43] B. Plale, BD Hubs: Collaborative Proposal: West: Accelerating the Big Data Innovation Ecosystem: NSF Award #1916481. Retrieved November 26, 2019, from https://www.nsf.gov/awardsearch/showAward?AWD_ID=1916481&HistoricalAwards=false, 2019.

[44] B.A. Plale and A.M. Suarez, Big Data Regional Innovation Hubs. Retrieved November 26, 2019, from https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505185, 2015.

[45] Schmidt Futures. Schmidt Futures. Retrieved December 1, 2019, from https://schmidtfutures.com/our-work/technology-society/, 2019.

[46] R.F. Service, Department of Energy plans major AI push to speed scientific discoveries. Retrieved November 26, 2019, from Science website https://www.sciencemag.org/news/2019/10/department-energy-plans-major-ai-push-speed-scientific-discoveries, 2019.

[47] C. Tenopir, S. Allard, K. Douglass, A.U. Aydinoglu, L. Wu, E. Read and M. Frame, Data sharing by scientists: Practices and perceptions, *PLoS ONE* **6**(6) (2011), e21101. doi:10.1371/journal.pone.0021101.

[48] C. Tenopir, L. Christian, S. Allard and J. Borycz, Research data sharing: Practices and attitudes of geophysicists, *Earth and Space Science* **5**(12) (2018), 891–902. doi:10.1029/2018EA000461.

[49] C. Tenopir, E.D. Dalton, S. Allard, M. Frame, I. Pjesivac, B. Birch and K. Dorsett, Changes in data sharing and data reuse practices and perceptions among scientists worldwide, *PLOS ONE* **10**(8) (2015), e0134826. doi:10.1371/journal.pone.0134826.

[50] top500. TOP500 Supercomputer Sites. Retrieved November 26, 2019, from https://www.top500.org/lists/2019/06/, 2019.

[51] X. Wang, S. Xu, L. Peng, Z. Wang, C. Wang, C. Zhang and X. Wang, Exploring scientists' working timetable: Do scientists often work overtime? *Journal of Informetrics* **6**(4) (2012), 655–660. doi:10.1016/j.joi.2012.07.003.

[52] M.D. Wilkinson, M. Dumontier, Ij.J. Aalbersberg, G. Appleton, M. Axton, A. Baak and B. Mons, The FAIR Guiding Principles for scientific data management and stewardship, *Scientific Data* **3**(1) (2016), 160018. doi:10.1038/sdata.2016.18.