

Geoscience data publication: Practices and perspectives on enabling the FAIR guiding principles

Danie Kinkade   | Adam Shepherd

Biological and Chemical Oceanography
Data Management Office, Woods Hole
Oceanographic Institution, Woods Hole,
MA, USA

Correspondence

Danie Kinkade, Biological and Chemical
Oceanography Data Management Office,
Woods Hole Oceanographic Institution,
Woods Hole, MA, USA.
Email: dkinkade@whoi.edu

Funding information

Division of Ocean Sciences, Grant/Award
Number: 1924618

Abstract

Introduced in 2016, the FAIR Guiding Principles endeavour to significantly improve the process of today's data-driven research. The Principles present a concise set of fundamental concepts that can facilitate the findability, accessibility, interoperability and reuse (FAIR) of digital research objects by both machines and human beings. The emergence of FAIR has initiated a flurry of activity within the broader data publication community, yet the principles are still not fully understood by many community stakeholders. This has led to challenges such as misinterpretation and co-opted use, along with persistent gaps in current data publication culture, practices and infrastructure that need to be addressed to achieve a FAIR data end-state. This paper presents an overview of the practices and perspectives related to the FAIR Principles within the Geosciences and offers discussion on the value of the principles in the larger context of what they are trying to achieve. The authors of this article recommend using the principles as a tool to bring awareness to the types of actions that can improve the practice of data publication to meet the needs of all data consumers. FAIR Guiding Principles should be interpreted as an aspirational guide to focus behaviours that lead towards a more FAIR data environment. The intentional discussions and incremental changes that bring us closer to these aspirations provide the best value to our community as we build the capacity that will support and facilitate new discovery of earth systems.

KEYWORDS

data infrastructure, domain data repositories, FAIR Guiding Principles, geosciences, research data publication

1 | INTRODUCTION

Over a century ago, the scientific research community first acknowledged the importance and value data hold beyond their original, intended use (Cajal, 1999). Geoscience data

hold particular value, representing snapshots of Earth systems that are unique in both space and time, and as such are often irreplaceable. These data serve as records of the past and present, necessary to understand current, and predict future states, rates and processes of global systems. But the

Dataset details: No data were collected or used during the course of preparing this manuscript.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2021 The Authors. *Geoscience Data Journal* published by Royal Meteorological Society and John Wiley & Sons Ltd.

nature of geoscience data has evolved and the size, speed and variety in which earth observations are produced today pose significant challenges to data sharing and reuse.

In long-tail geoscience disciplines, it is no longer practical (nor in some cases feasible) for researchers to store and work with data locally using a laptop computer. Because many of today's geoscience researchers need to work with data that span multiple domains, the ability to discover, access and integrate data, often at scale, is crucial in their interdisciplinary research workflows. Likewise, there is great potential in artificial intelligence and machine learning that is presently hamstrung by the heterogeneity of today's data management practices. Yet the culture of sharing and managing data sufficiently to enable its reuse has been slow to gain footing in many scientific domains, even though scientists generally agree these practices are important and are willing to reuse data created by others (Enke et al., 2012; Roche et al., 2015; Tenopir et al., 2011; Tenopir et al., 2018).

Realizing the potential value of shared data to the research process, data publication stakeholders are driving change in the culture of data sharing, from funders wishing to maximize investment through data reuse, to the journal reviewer requiring access to evaluate scientific results (Costas et al., 2013; Division of Ocean Sciences (OCE) Sample and Data Policy(Nsf17037) | NSF - National Science Foundation; , n.d.; Guidelines to the Rules on Open Access to Scientific Publications & Open Access to Research Data in Horizon, 2020, 2017; Holdren, 2013; Mayernik, 2012, 2017). But simply sharing project data via a laboratory website is not enough to enable effective and efficient reuse. This endeavour requires a host of activities occurring at specific stages throughout the research data lifecycle that facilitates data discoverability, access and analysis. It requires a robust infrastructure, a data-savvy research workforce, and skilled data professionals. Collectively, these components form a data publication ecosystem, the maturity of which may vary widely from one research community to the next, in part because the details of resources and activities necessary to enable effective data discovery and reuse are not well defined. Austin et al., (2015) describe the research data publication workflow in a generalized reference model with key components; however, in practice there are many nuances not captured in this model due to domain specificity of both geoscience researcher and repository workflows.

2 | ENTER THE FAIR GUIDING PRINCIPLES

In 2016, Wilkinson et al. published *The FAIR Guiding Principles for scientific data management and stewardship* in an attempt to provide both clarity and guidance on the types of activities that would increase the reuse of digital

BOX 1 The FAIR guiding principles (Wilkinson et al., 2016)

To be Findable:

- F1. (meta)data are assigned a globally unique and persistent identifier
- F2. data are described with rich metadata (defined by R1 below)
- F3. metadata clearly and explicitly include the identifier of the data it describes
- F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:

- A1. (meta)data are retrievable by their identifier using a standardized communication protocol
 - A1.1 the protocol is open, free and universally implementable
 - A1.2 the protocol allows for an authentication and authorization procedure, where necessary
- A2. metadata are accessible, even when the data are no longer available

To be Interoperable:

- I1. (meta)data use a formal, accessible, shared and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles
- I3. (meta)data include qualified references to other (meta) data

To be Reusable:

- 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

research objects (broadly described as data, tools, algorithms and related workflows) (Box 1). Because today's scientific workflows are intrinsically dependent upon computational resources to address the scale associated with contemporary research, the focus of the principles is intentionally on machine-actionable strategies (human-readability being implicit).

There are some key aspects to note that impact data reuse, but are not directly addressed by the principles. The community adopting FAIR requests that data be as open as possible, where open means free from use restrictions, fees or embargos. However, FAIR data can be restricted, often for legal, moral or ethical reasons such as safeguarding patient personal identifying information (PII), fishery-dependent or

endangered species data. In these cases, FAIR deems that metadata should be available for discovery and access and explicitly describe the conditions under which data may be accessed and used. In addition, the principles do not cover the quality of observational values themselves. A data set may adhere to the principles, but be suspect in quality or state of completeness to enable its reuse in research. Likewise, the utility or fitness for purpose of data will differ among user audiences and specific research questions. The principles do not consider aspects of long-term preservation or curation of data. They require that metadata be accessible even if data are not available; however, this fact does little for machines to operate on the observations themselves if only metadata remains. The Principles do not recommend or endorse specific implementation strategies, so standards and technologies used to create FAIR data will undoubtedly differ across domains. This leaves the process of mapping structured information a remaining necessity for cross-domain data integration. In instances where no suitable FAIR standard or vocabulary is present in a research community, the suggested solutions are to extend a particular resource or create a new FAIR one; however, each of these activities is non-trivial.

Although Wilkinson et al. (2016) acknowledge that the desired end-state—a research environment where data may be discovered, accessed and reused, fully unaided by human intervention—may seldom occur in practice, the objective of the principles is to advance the flow of data along the continuum towards that goal. Across several domains, the F and A in FAIR have been more readily achieved than either I or R. It is also important to acknowledge the distinction between data and software with respect to FAIR principles, the latter having additional qualities that are not addressed by the FAIR Data Principles, such as executability. Presently, there is a concerted effort to create FAIR principles specific to research software; however in this paper, the authors focus on data and do not address the nature of rapidly evolving software principles.

3 | WHAT DOES FAIR MEAN FOR GEOSCIENCE DATA STAKEHOLDERS?

The emergence of FAIR has initiated much activity focused on data sharing, although many efforts and activities to move data towards a FAIR state were underway long before the publication of the principles. The FAIR acronym has provided the broader research data stakeholder community with a means of addressing them collectively. This, by and large, is beneficial to science.

The implementation of the FAIR Principles requires all stakeholders in the research ecosystem, since no single stakeholder possesses the skills and resources to implement the

FAIR Principles themselves. Endeavouring to achieve a FAIR data end-state within the geosciences necessarily requires capacity building throughout the research and data lifecycle. Skills, services and infrastructure are needed to realize end-to-end workflows that result in FAIR data. Each stakeholder in that workflow has a role to play in ensuring any given digital resource resulting from geoscience research is reusable. This includes the data originator and the data curator, in addition to the broader research and information science communities, who establish best practices and standards.

Implementation has begun in several countries, to educate, and develop capacity that will support a FAIR data ecosystem. Africa, Japan and Australia have committed to developing data infrastructure in support of FAIR and Open Data (Digital Science et al., 2018). In Europe especially, considerable resources have been made available to support efforts aimed at advancing FAIR activities from both top-down and bottom-up approaches. Efforts such as GO FAIR (GO FAIR, n.d.), a collaboration between France, Germany and the Netherlands, and FAIRsFAIR (FAIRsFAIR, n.d.), a project funded by the European Commission, aim to advance FAIR capacity through policy, education and service development. Within the United States, the Coalition for Publishing Data in the Earth and Space Sciences (COPDESS) and the American Geophysical Union's Enabling FAIR Data project collaborate to educate and bring together various data publication stakeholders in an effort to build consensus, coordinate workflows and leverage existing resources that help researchers and repositories work towards achieving FAIR together (Stall et al., 2018; Stall et al., 2019). Professional organizations and societies such as the International Science Council's Committee on Data (CODATA) (CODATA, n.d.), American and European Geophysical Unions (AGU, EGU), the Research Data Alliance (RDA, Berman, 2019), the Future of Communication and e-Scholarship (FORCE11, n.d.), the Earth Science Information Partners (ESIP, Earth Science Information Partners, 2020) and others are shifting focus towards efforts that support the implementation of the FAIR Principles. Indeed, across the globe, there have been workshops and webinars, professional meeting sessions, Data FAIRs (meeting exhibits aimed at increasing FAIR awareness), symposiums and a host of other events aimed at promulgating and proselytizing the virtues of FAIR as it relates to broader Open Science processes.

Researchers are becoming increasingly aware that they must do more to prepare their data for sharing and reuse, whether enticed by the benefits of increased altmetrics through data publication citations, motivated by the funders' enforcement that federally funded research output be shared or directed by journals to ensure their data are accessible to reviewers. But so far, it would appear they are not completely sure what 'achieving' FAIR means for them in the broader context of its objectives. The annual report, *The State of Open*

Data 2018 (Digital Science et al., 2018), found that nearly 60% of researchers surveyed ($n = 748$) had never heard of the FAIR Guiding Principles. Of those that had, over 40% felt that they needed further clarification on its definitions. In the recently released 2020 report, the number of researchers who had never hear of FAIR decreased to just under 40% (Digital Science et al., 2020). However, there remains a large knowledge gap between hearing the term 'FAIR' and fully understanding the principles. A study by Bishop et al. (2019) looked at geoscience researcher (meta)data search and reuse habits in the context of the FAIR Principles and found that many did not understand the value of structured metadata, the meaning of use constraints or concept of controlled vocabularies.

Perhaps this is due to the fact that although data originators are realizing they must actively include data management in their research workflows, education on data management planning, good data hygiene and repository selection are still not well incorporated into today's science pedagogy. A study by Tenopir et al. (2018) found that less than 30% of geoscience respondents ($n = 1,372$) felt their organizations or projects provided training or assistance in data management best practices or metadata creation.

But, researchers and data originators play perhaps the biggest role in making data FAIR by providing a foundation on which to apply the principles. If data collection is poorly executed, or critical metadata not captured, then no subsequent effort downstream from data professionals and repositories can improve that state to enable the *Interoperability* and *Reusability* in FAIR. From a research project's planning stages, investigators should be considered the types of data they will produce. It is at this stage that they should be aware of their community's standards and best practices surrounding those data types. Are there well-known, or frequently used vocabularies with which they can describe their data? Likewise, for file formats and unit conventions, are there domain-specific repositories that their community routinely uses to share and discover data? Will any of the data need to be embargoed or restricted?

Ideally, these types of questions would be answered and documented in a data management plan (DMP). Oftentimes, funders have set guidelines or requirements on data sharing aspects such as embargo allowances or recommended repositories that should be incorporated into DMPs. Here, institutional libraries and repositories are excellent resources for data management-related strategies. For example, they can point researchers to online resources such as FAIRsharing (fairsharing.org) a registry of data standards and policies active in promulgating FAIR resources (Sansone et al., 2019) and the Registry of Research Repositories (Re3data.org) to aid in data management planning. But, so far, it is unclear whether data originators perceive value in collaborating with repositories for their initial project planning.

Researchers should also determine the key pieces of metadata that need to be captured prior to data collection. Geoscience observations are unique in time and space, and therefore cannot be resampled. So, it is critical that observations have associated sample metadata at a minimum (i.e. date, time, location and possibly depth or altitude). Metadata critical for the R in FAIR (i.e., R1, R1.3) include documentation and/or links to methodologies, sampling protocols, instrumentation and calibrations, quality control and assurance information, associated reports, related publications and related data sets or accession numbers.

In some geoscience communities, the relationship between individual scientists, their research community and its designated domain repository is tightly coupled, facilitating good practices. For example, in oceanography, the International GEOTRACES Programme (www.geotraces.org), which aims to understand trace element isotope distribution and related biogeochemical cycling in global oceans, established a dedicated International Data Management Committee. This body provides recommendations that drive a comprehensive data management workflow that includes individual researchers, the International Data Centre and supporting National Data Assembly Centers. GEOTRACES researchers must comply with the programme's data policy which outlines specific metadata requirements such as data and unit nomenclature using community accepted vocabularies. This helps facilitate integration of all data into a collective product.

The relatively new and rapidly evolving ocean proteomic community is another example where researchers are coming together to develop best practices and a common data model that would allow their data to be easily discovered and analysed across multiple laboratories over time (Saito et al., 2019).

These programmes build on lessons learned from previous decadal-scale research efforts in oceanography (e.g. the World Ocean Circulation Experiment (WOCE, Woods, 1985) and the Joint Global Ocean Flux Study (JGOFS, Hanson et al., 2000)). Such initiatives were international and collaborative in scope and had organized data management coordination offices to enable data integration and synthesis activities. This aligns with evidence that data sharing attitudes and practices may be more established in domains where data do not pertain to human subjects, sharing has a long history, metadata standards are established, and large-scale instrumentation is shared among data collectors (i.e. within oceanography, research vessels and their associated water sampling instrumentation are often shared across projects) (Tenopir et al., 2018).

What about researchers who do not have an organized or cohesive community endeavouring to develop or promulgate standards or best practices? Although no one is expecting individual scientists, themselves, to make their data FAIR compliant, they should realize that the decisions and strategies

they use at the beginning of their project help enable their data to become FAIR. In the words of Lambert Heller ‘As a scientist, you should treat your data like a love letter to your future self’ (Brock, 2020). Proactive geoscience researchers desiring to improve their data management skills have several resources available to them, including those developed within the Earth and space science information communities, such as ESIP’s Data Management Training Clearinghouse (<https://dmtclearinghouse.esipfed.org>) and DataONE’s Data Management Training Short Course (<https://www.dataone.org/training>), in addition to more formal resources such as the Data Carpentries (<https://carpentries.org>) and the FORCE11 Scholarly Communications Institute (<https://www.force11.org/fsci/>). Registries such as FAIRsharing and Re3data can help geoscientists locate suitable repositories for sharing their data. However, the decisions necessary to select an appropriate facility are often complex and need to consider emerging funder and publisher requirements (Enabling FAIR Data Community et al., 2018). More socialization of these available resources would undoubtedly improve this situation, as would efforts to more broadly increase data skill capacity among geoscience researchers.

3.1 | What about data repositories?

Beyond the role of the data originator ensuring observations and metadata are captured in a way that facilitates reuse, the responsibility of implementing many of the FAIR Principles falls primarily on data professionals and supporting repositories. Wilkinson et al. (2018) described the need for highly specialized domain repositories the likes of those in the life and space sciences, because these repositories bring subject matter expertise to the data curation process. Domain repositories combine scientific knowledge with information management skills, work closely with their research communities to apply quality controls, create and curate robust discovery- and use-level metadata and document provenance, thereby increasing data reusability. In addition, they apply harmonization techniques to data and metadata that increase interoperability across scientific domains. They employ persistent identifiers for metadata and related information, and standardized formats for representing metadata and data; they facilitate the application of appropriate licences and make connections to qualified external resources for additional interpretation and context for reuse. They are, without question, critical to the creation of a FAIR Data ecosystem.

It is also worth noting that although the FAIR Principles were first published in 2016, the development of infrastructure aimed at its objectives has been occurring in the data management and curation communities for decades. Metadata standards such as Dublin Core (Weibel et al., 1998) and ISO 19139 (<https://www.iso.org/standard/32557.html>)

have been used by librarians and data managers since 1998 and 2007, respectively. Likewise in the geosciences, meta-data standards such as Ecological Markup Language (EML, <https://eml.ecoinformatics.org>) and ISO 19139 extensions such as ISO 19115-2 ensure machine-actionable discovery is possible through shared metadata schemas. Yet in spite of broad adoption, these standards suffer from a variety of (mis)uses and non-conformance to the specifications making findability and interoperability difficult for content aggregators (Díaz et al., 2010; O’Dea et al., 2010); a fact that highlights persistent challenges when aspiring to the FAIR Principles.

There are 743 geoscience repositories identified in the Re3data registry. The authors of this article are principal investigators for The Biological and Chemical Oceanography Data Management Office (BCO-DMO), a domain-specific geoscience repository funded primarily by the US National Science Foundation. BCO-DMO works directly with the data producers and investigators conducting ocean ecosystem research, from the point of data management planning through data collection and publication, to ensure data are well-described and as clean and accurate as possible. In the process, the office helps researchers fulfil their funding requirements, educates them in basic data skills and contributes their data to a growing, searchable and freely accessible catalog. In order to provide domain expertise at scale, the office endeavours to automate as much as possible and employs an adaptive approach to its infrastructure by managing resources using W3C Recommended Best Practices for Data on the Web (<https://www.w3.org/TR/dwbp/>). This approach was largely driven by the rapidly changing needs of the user community for greater information to be incorporated into metadata (for data set discovery and reuse). Since these needs were evolving faster than updates were occurring to community metadata standards, the office formalized its own internal metadata model into an ontology. With this ontology, better described as a BCO-DMO-specific data model, information is stored, queried and accessed from a central source known as a knowledge graph. Graph structures are easy to extend over time, and the office can quickly respond to user needs by first updating the data model to include new elements, then populating the knowledge graph with the new element information. Metadata records are then produced in a number of standard formats using information from the knowledge graph. This strategy helps the repository maintain the distinction between what it knows about research data and what a specific metadata standard requires to enable discovery. Such strategies were developed before FAIR was coined, and indeed, there are many other domain-specific repositories such as BCO-DMO that rely on technological and subject matter expertise to apply the FAIR Principles to their data catalogs.

But the FAIR principle authors also recognized a surge of generalist repositories in recent years, from small institutional facilities to those international in scope. This

growth poses a challenge in the heterogeneity and potential inconsistency of data types likely to be encountered by machines, to say nothing of the need to distinguish authoritative sources among multiple copies of a data set residing in different catalogs. The authors go on to argue this fact increases the need for generalized approaches by repositories. However, the intentional generality of the principles makes implementation as heterogeneous as the repository landscape, itself. Superimpose domain-specific practices such as metadata standards and file conventions, a general disagreement on what constitutes a minimum amount of metadata across the geosciences and arriving at one-size-fits-all approaches seem a continuing and ever-impossible to reach challenge.

Along with the growing number of repositories (both generalist and domain-specific), is the increasing diversity of their capacity and maturity in terms of data management expertise and infrastructure. Different data facilities possess unique mandates, missions and target communities, and their level of curation and the strategies they employ vary widely. So much so, that it is difficult to determine which repositories are capable of fully implementing the FAIR Principles. This poses problems for data stakeholders seeking a suitable repository for sharing data and ensuring its access and reuse.

3.2 | What about other stakeholders?

There are additional stakeholders in the data publication workflow who, through their own drivers, are applying additional pressures to the system. For example, over the past decade the United States and European Union governments established policy and recommendations aimed at increasing data sharing in an effort to maximize their investments through the reuse of Federally funded research output (e.g. US Office of Science and Technology Policy Memo *Increasing Access to Results of Federally Funded Science Research*, Horizon2020 Data guidelines, and the European Open Science Cloud). In the United States, impacts of such policy have cascaded through federal and private funders as they establish or update data sharing requirements (e.g. the US National Science Foundation (NSF), National Oceanographic and Atmospheric Administration (NOAA), the Bill and Melinda Gates Foundation and the Gordon and Betty Moore Foundation). Publishers, such as Elsevier, Springer Nature and Wiley, have also developed policies and guidelines for reviewers and authors, ensuring that data supporting scholarly publications are shared in a trustworthy and public repository. However in most journals, the focus is on only the data contributing to the publication, and therefore, remaining project outputs are not considered and may go unshared.

These varied stakeholder drivers have resulted in increased pressure for geoscience data repositories to accept

and curate data resulting from highly heterogeneous research. In some cases, further taxing a data curation system that has grown in an organic fashion to meet the needs of various research communities.

Professional societies and organizations such as the AGU, CODATA, RDA, ESIP and FORCE11 are playing a key role in supporting development of e-infrastructure from the bottom-up, while recognizing and providing connections to global governmental policies and top-down drivers. Serving as sources and sustaining bodies for best practices and guidelines, they facilitate adoption of many technologies and practices that enable repositories to implement the FAIR Principles. These organizations will undoubtedly continue to be important and necessary assets on the journey to a FAIR data ecosystem.

4 | METRICS FOR FAIRness... ARE THEY FAIR?

Although the FAIR Principles have proven to be an impressive communication tool, presenting simple actions or attributes that can improve data reuse, they have quickly been leveraged to develop metrics and assessment frameworks aimed at evaluating the implementation of those same actions and attributes. Increasingly, compliance with the fifteen sub-principles is being described as determining 'FAIRness', not just of a data set, but also of the agents applying the principles to data sets, namely repositories. Therefore, not surprisingly, attention is being turned to repository practices and operations.

In 2017, Dunning et al. looked at the effectiveness and relevance of the principles as a metric for repository capacity of FAIR implementation. The study assessed the application of services and practices of over 40 repositories in the Netherlands to determine whether a repository was implementing the FAIR sub-principles. They found less than 50% adhered to the Findable sub-principles, and although nearly 100% compliance was found for most of the Accessible sub-principles, hardly any of the repositories demonstrated metadata persistence in the absence of data availability. While interoperability overall was achieved, reusability was difficult for repositories to demonstrate. In general, the project found some of the sub-principles to be vague (e.g. F2) and redundant (e.g. F2 and R1), some to be subjective (e.g. R1 and R3), and others to be recursive (e.g. I2), or dependent upon repository policy, which may be dependent upon the needs of the individual communities they serve (Dunning et al., 2017).

A year later, the original authors of the FAIR Principles acknowledged that they were being interpreted in a variety of ways. Coincident was the desire on behalf of nearly every stakeholder in the data publication ecosystem to evaluate the application of the principles. So, in collaboration with key

FAIR stakeholders (e.g. publishers and policymakers), a framework for assessing FAIRness was created (Wilkinson et al., 2018) that consists of a template and subsequent exemplar FAIR maturity indicators. Like the principles themselves, the metrics focused on FAIRness for machines. But the aspirational intent and ambiguity of the principles that led to their misinterpretation also make any assessment difficult. The manner in which they can be applied will vary across communities, and what one stakeholder community finds necessary to demonstrate FAIRness may not resonate with another. No consideration was given to long-term curation, and a governance model for the framework was not provided. Additionally, although each metric refers to the quality of the data in question, the evidence needed to demonstrate a particular metric falls on the responsibility of the digital resource provider or repository. Repositories must therefore provide machine accessible documents as evidence of authorization procedures, persistent identifier strategy, etc. As in the Dunning study, the Wilkinson *et al.* metrics for FAIRness of digital research objects effectively become an assessment on repository practices.

These examples are not the only efforts attempting to leverage the principles as a metric; funders and publishers are also weighing in repository capabilities to implement FAIR. In January of this year, the US Office of Science and Technology Policy issued a request for public comment on a draft document for *Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research* (85 FR 12949). The draft presents a series of proposed characteristics for all repositories anchored in the FAIR Principles. The US National Institutes of Health (NIH) Office of Director has already incorporated the characteristics as supplemental information to its data management and sharing policy (<https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-016.html>). Although intended to improve consistency across U.S. agency practices and provide guidelines on data preservation, the document alludes to its potential use as an assessment tool. Commercial Publishers produced a similar guidance document to determine an appropriate repository for data deposition (Sansone et al., 2020), with the driver being preservation of the data necessary for the review of publication findings. The document aims to ease the decision-making process of authors, consolidate the multitude of individual publisher policies and inform repositories of publisher views on desired curation features.

These recent endeavours are similar in intent and have heavy overlap with already existing evaluation and certification frameworks to assess the quality and trustworthiness of repository operations (such as the ISO16363 Standard for Trusted Digital Repositories, the NESTOR Seal for Trustworthy Digital Archives (nestor, n.d.), and the CoreTrustSeal (Dillo & Leeuw, 2018)), and yet are distinct

efforts, being driven by each group's unique needs comprising larger supersets of requirements.

This broader activity to leverage FAIR as an assessment tool points to the principles themselves being far more impactful than the details of their implementation. Measuring FAIRness is only valuable across data curation operations that are driven by the same goal. To measure FAIRness across varied goals and missions does not produce anything truly useful for consumers. More importantly, this practice is putting new and rapidly evolving pressures on geoscience repositories, as they try to meet each stakeholder's requirements and expectations. Ironically, one intent of Sansone et al. (2020) is to increase efficiency for data repositories dealing with multiple publishers; however, the very proliferation of 'repository criteria' documents is having just the opposite impact as data facilities need to address new expectations on several fronts. Performing self-assessments, undertaking a formal certification or demonstrating FAIRness come at a cost in both repository time and budget. Many facilities have evolved over decades to serve particular research disciplines and operate primarily on grant-based support, so their ability to pivot and adapt to new stakeholder demands is constrained by grant cycles or prescribed missions that leave little room for infrastructure innovation. What are the possible downstream effects on a repository and its supported research community if it fails to meet any of various FAIR-related assessment criteria? For example, when an author seeking to publish their research results has a funder requirement to use a designated repository that is in conflict with the nascent journal criteria or when results of FAIRness criteria are used to make funding decisions on repositories that serve specific research communities, yet do not meet the FAIRness bar.

It remains to be seen if data publication stakeholders at large perceive value in building capacity in the existing geoscience research data infrastructure to achieve a more FAIR data landscape. Likewise, if these stakeholders can converge on a single assessment tool (such as the CoreTrustSeal) to demonstrate the robustness of repository capacity to enable the FAIR Principles.

5 | ASPIRING TO BE FAIR

In an attempt to provide a greater detail and rationale for the FAIR Principles, Jacobsen et al. (2019) present interpretation and implementation considerations of each sub-principle, promoting the notion of a necessary community convergence on strategies and technologies to successfully achieve an Internet of FAIR objects. However, the sheer multitude of implementation choices for some principles (e.g. metadata standards), juxtaposed with a dearth of choices for others, highlights the ongoing challenges that face many geoscience communities.

Given all of the current attention and activities surrounding the FAIR Guiding Principles, is it possible to achieve a data publication environment suitable for both humans and machines, where geoscience data are findable, accessible, interoperable and reusable? The desire to realize the FAIR Principles and monitor progress towards them may depend on one's perceived value of a FAIR Data end-state and the intent for applying the principles. In the life sciences and more specifically biopharmaceutical research and development entities, there is potential to realize significant return on investment with the achievement of FAIR via the exploitation of vast amounts of data using tools such as machine learning and artificial intelligence. Long-term impacts allude to the potential for faster treatment discoveries and increased efficiencies in the R&D pipelines (Wise et al., 2019). These benefits hinge upon the implementation of the FAIR sub-principles and gauging progress towards that goal.

The advantages of exploiting big data to solve pressing challenges transcend many scientific domains, and geoscience researchers are increasingly faced with the need to find, access and analyse large amounts of highly heterogeneous data to answer complex questions about Earth's systems. As data volumes continue to grow, researchers will need to rely on more automated processes versus reliance on bespoke analytical workflows. And scientists are beginning to share their experiences and the value of practicing Open Science, where data, analysis tools and workflows are findable, freely accessible and reusable (Lowndes et al., 2017; McKiernan et al., 2016). An excellent example of efforts aligned with FAIR's goal of machine actionability is those of the newly launched NSF-supported AI Institute in Weather, Climate and Coastal Oceanography (AI2ES, n.d.). The institute is focused on improving weather models through research on AI algorithms to leverage the massive amounts of data needed to address pressing atmospheric and oceanographic challenges and increase the accuracy and reliability of weather predictions.

Considering that the principles themselves are conceptual by design, then attempts to assess their implementation will remain a potentially misguided challenge. It is therefore perhaps better to view the principles as inspiration for data publishers, to challenge assumptions and direct focus to the roles and needs of all data stakeholders (including machine consumers). The FAIR Principles should not be used as a metric to evaluate the state of a given data set or piece of infrastructure. Instead, they should serve as a guide to move us closer to FAIRness along a continuum of FAIR activities.

Taking this approach, the question then becomes one of which investments we can make that will move the needle towards more FAIR data. Although researcher attitudes are changing towards data sharing (the first step in making data FAIR), researcher confidence is low (Barone et al., 2017), incentives have been slow to take hold, and curriculum in

the geosciences at the undergraduate and graduate levels is nascent at best. As explained in Lowndes et al. (2017), researchers need to be not only socialized to the tools and strategies that can help them manage and analyse data better, but also they need to feel proficient in their use as well. Fostering better relations between data scientists and repositories, and researchers can help here, and collaborations to cross pollinate geoscience with information science students can lead to a more data-savvy workforce.

Consensus among other stakeholders (e.g. funders and publishers) on their own roles, needs and perspectives on the value of curation infrastructure can help move subsequent data-related activities in a positive direction along the FAIR continuum. Ensuring that repository managers are present in strategy discussions that directly impact data curation facilities will be crucial to set realistic expectations for near and long-term goals and objectives. Providing novel opportunities for innovation and collaboration among data repositories in order to bring basic curation practices up to a level that increases the FAIRness of geoscience data will be necessary. In the United States, the Council of Data Facilities (CDF, Earthcube, n.d.) serves as a coordinating body for geoscience repositories and aims to foster collaborations, promote the use of standards and explore shared infrastructure opportunities. Creating similar forums for coordination of data curation infrastructure will help create a more federated data framework that can better support machine agents finding, accessing and ultimately reusing geoscience data.

These activities are all a departure from our current practices and need financial investment and nurturing to succeed. It is hopeful that projects such as FAIRsFAIR, GO FAIR and Enabling FAIR Data, along with consortia and organizations such as RDA, ESIP and FORCE11, are endeavouring to train data stakeholders in good data management practices, facilitate standards development and adoption, and support repositories in developing capacity that will bring the data they curate closer to FAIR. Sustaining that capacity will undoubtedly take another culture shift towards a better understanding and appreciation of the role scientific data curation infrastructure plays in enabling data to become FAIR.

6 | CONCLUSIONS

At its highest level, the FAIR Data Principles are designed to improve aspects of data that enable its reuse via domain agnostic practices that recognize the potential in machine-actionable workflows. The acronym has been effective at socializing some basic activities that can be applied to increase the findability, accessibility, interoperability and reusability of a broad array of digital objects. But their vague definitions lead to differences in interpretation and subsequent implementation, and data considered FAIR within one science

domain and may still not be interoperable or reusable by another.

In the rush to promote FAIR among data publication stakeholders, the principles have been repurposed as a metric to demonstrate compliance and implementation; whether as a proxy to evaluate the capabilities of a repository or to determine the FAIRness of a digital object. An act poses challenges as we struggle to figure out the value of the principles in the larger picture of what they are trying to accomplish.

Perhaps our best approach is to use the FAIR Principles as a tool to bring awareness to the types of actions that can help the practice of data publication better meet the needs of all data consumers, human and machine. They should be interpreted as aspirational guidelines to focus behaviours that lead towards a more FAIR data environment rather than as a specific metric for evaluation. Activities and resources necessary to apply the principles will undoubtedly rely on culture change and capacity building. But if achieved, great scientific rewards may be realized by unlocking the potential held in the vast amounts of geoscience data being produced today.

ACKNOWLEDGEMENTS

The writing of this article was supported by the NSF, grant no. 1924618. The authors thank Shelley Stall, Erin Robinson, Lisa Raymond and Jaci Saunders for their domain expertise and comments during the drafting of this paper.


CONFLICT OF INTEREST

The authors declare that there are no conflicts of interest that could be perceived as prejudicing the impartiality of this article's content.

ORCID

Danie Kinkade  <https://orcid.org/0000-0002-1134-7347>

TWITTER

Danie Kinkade @  BCODMO

REFERENCES

- AI2ES (n.d.). NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES). Available at: <https://www.ai2es.org/> [Accessed 24 August 2020].
- Austin, C.C., Bloom, T., Dallmeier-Tiessen, S., Khodiyar, V., Murphy, F., Nurnberger, A. et al. (2015) *Key components of data publishing: Using current best practices to develop a reference model for data publishing*. Zenodo. <https://doi.org/10.5281/zenodo.34542>
- Barone, L., Williams, J. and Micklos, D. (2017) Unmet needs for analyzing biological big data: A survey of 704 NSF principal investigators. *PLOS Computational Biology*, 13(10), e1005755. <https://doi.org/10.1371/journal.pcbi.1005755>
- Berman, F. (2019) The research data alliance -- the first five years. Available from: <https://www.rd-alliance.org/sites/default/files/attachment/RDA%20RETROSPECTIVE%20FINAL%20-%20HDSR.pdf>
- Bishop, B.W., Hank, C., Webster, J. and Howard, R. (2019) Scientists' data discovery and reuse behavior: (Meta)data fitness for use and the FAIR data principles. *Proceedings of the Association for Information Science and Technology*, 56(1), 21–31. <https://doi.org/10.1002/pr2.4>
- Brock, J. (2020, February 11). "A love letter to your future self": what scientists need to know about FAIR data [Blog]. Nature Index. Available from: <https://www.natureindex.com/news-blog/what-scientists-need-to-know-about-fair-data> [Accessed 18 August 2020].
- Cajal, S.R. (1999) *Advice for a Young Investigator*. Cambridge, MA: MIT Press.
- CODATA. (n.d.). *About CODATA*. Available from: <https://codata.org/about-codata/> [Accessed 18 August 2020].
- Costas, R., Meijer, I., Zahedi, Z. & Wouters, P. (2013) *The value of research data—Metrics for datasets from a cultural and technical point of view. A Knowledge Exchange Report (A Knowledge Exchange Report)*. Available from: <https://www.researchgate.net/deref/http%3A%2F%2Fwww.knowledge-exchange.info%2Fdatametrics>.
- Data Science., Hahnel, M., Fane, B., Treadway, J., Baynes, G., Wilkinson, R. et al. (2018) The State of Open Data Report 2018 [Report]. *Digital Science*, <https://doi.org/10.6084/m9.figshare.7195058.v2>
- Díaz, P., Masó, J. & Guimet, J. (2010) *Comparative quality assessment of metadata: Two regional SDI case studies*. Proc. INSPIRE Conf. Krakow, Poland. 2010.
- Digital Science, Hahnel, M., McIntoshborrelli, L., Hyndman, A., Baynes, G., Crosas, M. et al. (2020) The State of Open Data 2020 [Report]. *Digital Science*. <https://doi.org/10.6084/m9.figshare.13227875.v2>
- Dillo, I. and Leeuw, L. (2018) CoreTrustSeal. *Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare*, 71(1), 162–170. <https://doi.org/10.31263/voebm.v71i1.1981>
- Division of Ocean Sciences (OCE) Sample and Data Policy (nsf17037) | NSF - National Science Foundation (n.d.) Available from: <https://www.nsf.gov/pubs/2017/nsf17037/nsf17037.jsp>
- Dunning, A., de Smaele, M. & Böhmer, J. (2017) Are the FAIR data principles fair? *International Journal of Digital Curation*, 12(2), 177–195. <https://doi.org/10.2218/ijdc.v12i2.567>
- Earth Science Information Partners. (2020) *Earth Science Information Partners (ESIP) Annual Report 2020 (Version 1)*. ESIP. <https://doi.org/10.6084/m9.figshare.13490274.v1>
- EarthCube. (n.d.). *Council of data facilities*. Available from: <https://www.earthcube.org/council-of-data-facilities>. [Accessed 24 August 2020].
- Enabling FAIR Data Community, Duerr, R., Kinkade, D., Witt, M. & Yarmey, L. (2018) Data repository selection decision tree for researchers in the earth. *Space, and Environmental Sciences*. <https://doi.org/10.5281/zenodo.1475430>
- Enke, N., Thessen, A., Bach, K., Bendix, J., Seeger, B. and Gemeinholzer, B. (2012) The user's view on biodiversity data sharing—Investigating facts of acceptance and requirements to realize a sustainable use of research data—. *Ecological Informatics*, 11, 25–33. <https://doi.org/10.1016/j.ecoinf.2012.03.004>
- FAIRsFAIR. (n.d.) *The project*. Available from: <https://www.fairsfair.eu/the-project>. [Accessed 17 August 2020].

- FORCE11. (n.d.) *FORCE11-RDA fairsharing working group*. Available from: <https://www.force11.org/group/biosharingwg>. [Accessed 18 August 2020].
- GO FAIR. (n.d.) *GO FAIR initiative*. Available from: <https://www.go-fair.org/go-fair-initiative/>. [Accessed 17 August 2020].
- Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020, v3.2 11. (2017) Available from: https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm
- Hanson, R.B., Ducklow, H.W. and Field, J.G. (eds). (2000) *The Changing Ocean Carbon Cycle: A Midterm Synthesis of The Joint Global Ocean Flux Study*. IGBP Book Series No. 5. Cambridge, UK: Cambridge University Press. pp. 520.
- Holdren, J.P. (2013) *Memorandum for the Heads of Executive Departments and Agencies: Increasing Access to the Results of Federally Funded Scientific Research*. Washington, DC: Office of Science and Technology Policy. Available from: <https://obamawhitehouse.archives.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research>.
- Jacobsen, A., de Miranda Azevedo, R., Juty, N., Batista, D., Coles, S., Cornet, R. et al. (2019) FAIR Principles: Interpretations and Implementation Considerations. *Data Intelligence*, 2(1–2), 10–29. https://doi.org/10.1162/dint_r_00024
- Lowndes, J.S.S., Best, B.D., Scarborough, C., Afflerbach, J.C., Frazier, M.R., O'Hara, C.C. et al. (2017) Our path to better science in less time using open data science tools. *Nature Ecology & Evolution*, 1(6), 1–7. <https://doi.org/10.1038/s41559-017-0160>
- Mayernik, M.S. (2012) Data citation initiatives and issues. *Bulletin of the American Society for Information Science and Technology*, 38(5), 23–28. <https://doi.org/10.1002/bult.2012.1720380508>
- Mayernik, M.S. (2017) Open data: Accountability and transparency. *Big Data & Society*, 4(2), 205395171771885. <https://doi.org/10.1177/2053951717718853>
- McKiernan, E.C., Bourne, P.E., Brown, C.T., Buck, S., Kenall, A., Lin, J. et al. (2016) How open science helps researchers succeed. *ELife*, 5, e16800. <https://doi.org/10.7554/eLife.16800>
- Nestor. (n.d.). *nestor Seal for Trustworthy Digital Archives*. Available from: https://www.langzeitarchivierung.de/Webs/nestor/EN/Zertifizierung/nestor_Siegel/siegel.html. [Accessed 24 August 2020].
- O'Dea, E., Dwyer, E., Cummins, V., Giménez, D.P.Í. & Dunne, D. (2010) Harmonising marine information exchange in Ireland. In: Green, D. (Eds.) *Coastal and Marine Geospatial Technologies. Coastal Systems and Continental Margins*, vol 13. Dordrecht: Springer. https://doi.org/10.1007/978-1-4020-9720-1_4
- Roche, D.G., Kruuk, L.E.B., Lanfear, R. & Binning, S.A. (2015) Public data archiving in ecology and evolution: How well are we doing? *PLOS Biology*, 13(11), e1002295. <https://doi.org/10.1371/journal.pbio.1002295>
- Saito, M.A., Bertrand, E.M., Duffy, M.E., Gaylord, D.A., Held, N.A., Herve, W.J. et al. (2019) Progress and challenges in ocean metaproteomics and proposed best practices for data sharing. *Journal of Proteome Research*, 18(4), 1461–1476. <https://doi.org/10.1021/acs.jproteome.8b00761>
- Sansone, S.-A., McQuilton, P., Cousijn, H., Cannon, M., Chan, W.M., Callaghan, S. et al. (2020) Data repository selection: Criteria that matter. *Zenodo*, <https://doi.org/10.5281/zenodo.4084763>
- Sansone, S.-A., McQuilton, P., Rocca-Serra, P., Gonzalez-Beltran, A., Izzo, M., Lister, A.L. et al. (2019) FAIRsharing as a community approach to standards, repositories and policies. *Nature Biotechnology*, 37(4), 358–367. <https://doi.org/10.1038/s41587-019-0080-8>
- Stall, S., Yarmey, L., Boehm, R., Cousijn, H., Cruse, P., Cutcher-Gershenfeld, J. et al. (2018) Advancing FAIR data in Earth, space, and environmental science. *Eos*, 99, <https://doi.org/10.1029/2018E0109301>
- Stall, S., Yarmey, L., Cutcher-Gershenfeld, J., Hanson, B., Lehnert, K., Nosek, B. et al. (2019) Make scientific data FAIR. *Nature*, 570(7759), 27–29. <https://doi.org/10.1038/d41586-019-01720-7>
- Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A.U., Wu, L., Read, E. et al. (2011) Data sharing by Scientists: Practices and perceptions. *PLoS One*, 6(6), e21101. <https://doi.org/10.1371/journal.pone.0021101>
- Tenopir, C., Christian, L., Allard, S. & Borycz, J. (2018) Research data sharing: Practices and attitudes of geophysicists. *Earth and Space Science*, 5(12), 891–902. <https://doi.org/10.1029/2018EA000461>
- Weibel, S., Kunze, J., Lagoze, C. & Wolf, M. (1998, September). *Dublin Core Metadata for Resource Discovery*. Available from: <https://www.hjp.at/doc/rfc/rfc2413.html>.
- Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A. et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018. <https://doi.org/10.1038/sdata.2016.18>
- Wilkinson, M.D., Sansone, S.-A., Schultes, E., Doorn, P., da Silva, B., Santos, L.O. et al. (2018) A design framework and exemplar metrics for FAIRness. *Scientific Data*, 5, 180118. <https://doi.org/10.1038/sdata.2018.118>
- Wise, J., de Barron, A.G., Splendiani, A., Balali-Mood, B., Vasant, D., Little, E. et al. (2019) Implementation and relevance of FAIR data principles in biopharmaceutical R&D. *Drug Discovery Today*, 24(4), 933–938. <https://doi.org/10.1016/j.drudis.2019.01.008>
- Woods, J. (1985) The world ocean circulation experiment. *Nature*, 314, 501–511. <https://doi.org/10.1038/314501a0>

How to cite this article: Kinkade D, Shepherd A. Geoscience data publication: Practices and perspectives on enabling the FAIR guiding principles. *Geosci Data J.* 2021;00:1–10. <https://doi.org/10.1002/gdj3.120>