ILLUSTRATION: THE PROJECT TWINS

HOW FAIR DATA ARE HELPING TO BUILD TRUST IN SCIENCE

The FAIR guidelines laid a solid foundation for data science. A decade on, researchers are thinking beyond them. **By Amanda Heidt**

If you build a data set and nobody can find it, is it useful? Not as much as it could be. With trust in science under siege from partisan actors and impartial pathogens, the accessibility and transparency of – and trust in – scientific information must be improved.

Enter the FAIR Data Principles. In 2014, scientists realized that data management and stewardship could benefit from a set of shared guidelines, and dozens of international researchers gathered to draft new recommendations. The resulting principles – which established that data should be findable, accessible, interoperable and reusable (FAIR) – were published ten years ago[1]. The original publication has around 16,000 citations, and governments, funders and publishers around the world now ask that data be hosted and shared in FAIR-compliant ways.

A decade on, however, even the founders acknowledge that the FAIR principles are an imperfect tool. Barend Mons, a molecular biologist at Leiden University in the Netherlands who conceived the initiative, says that FAIR was always meant to be a set of general principles, "and so, by definition, cannot address the specifics of every application". Fortunately, other researchers have taken the framework and extended it to cover the broader data ecosystem[2], including the algorithms, tools and workflows that drive contemporary research.

Making every discipline FAIR

At its core, FAIR is meant to ensure that data are produced, analysed, stored and shared in ways that promote transparency and reproducibility. "The more the data are understandable by people other than the creators, the more we are able to determine not only the trustworthiness of the data set itself, but also its alleged creators," says Mons.

The ideal data set should be properly documented, simple for both computers and people to find and use. It should also be easy to integrate with other data. To accomplish this, scientists must design workflows before data have been collected and create and maintain a detailed metadata file – an often overlooked component that contains contextual information about the data set, such as where and when it was created. The initiative also prioritizes data-management plans, including choosing appropriate licences and persistent identifiers – the unique labels ascribed to different resources – such that any information created during a project is findable and usable long after the research is over.

"It's a lot to think about, and I can see why it might seem really daunting for some scientists to consider," says Amelia Jiménez-Sánchez, a data-integrity researcher at the University of Barcelona in Spain. But FAIR is like cooking, she says: once you have the right ingredients – or familiarize yourself with FAIR practices – it becomes easier to make a meal. "Eventually, it just becomes a part of how you do your work."

Users have tailored those practices to their

disciplines. Carnegie Mellon University in Pittsburgh, Pennsylvania, has released FAIR guides for chemistry, mathematics, neuroscience and psychology. Other initiatives have focused on astronomy, materials science, genetics and single-cell genomics data. For fields without dedicated FAIR resources, researchers in the Netherlands have published 'ten simple rules' for kick-starting conversations about FAIR practices[3].

Recognizing that discipline-specific resources didn't exist in his field, Eliu Huerta, a theoretical physicist at the Argonne National Laboratory in Lemont, Illinois, began adapting FAIR principles for high-energy physics. Today, Huerta is part of a collaboration called FAIR4HEP, aimed at helping researchers in the field to improve their data-sharing practices. In 2022, he co-wrote a study evaluating data from the Large Hadron Collider at CERN, Europe's particle-physics laboratory near Geneva, Switzerland, for its 'FAIRness'[4]. Among other things, the study "provides a domain-agnostic, step-by-step set of checks to guide in the process of making a dataset FAIR", it says – a process the authors call FAIRification. The web-based FAIR Data Self-Assessment Tool from the Australian Research Data Commons, a company building research-data infrastructure in Melbourne, likewise offers "practical tips on how to enhance FAIRness" of your data.

Expanding beyond data

The FAIR guidelines also apply to software. The FAIR-USE4OS guidelines[5] extend FAIR principles to open-source software projects, for instance, and initiatives such as FAIR4RS focus on research software[6].

"Data are data, but there's also the entire system of infrastructure that is built around it to store, share and analyse that information, and those tools need to be fair and reproducible too," says Natalie Cooper, a macroecologist at the Natural History Museum in London.

Last year, Cooper edited a guide to reproducible code on behalf of the British Ecological Society that is rooted in FAIR principles. The code and data share many features, so a lot of the recommendations remain the same. But something she has found most helpful in her own work is code review, which Cooper now does before submitting anything for publication. During the review, colleagues exchange protocols, test them for reproducibility and suggest ways to improve efficiency. "You just make comments to each other, and hopefully you can improve each other's code," Cooper says. "It can be a really positive experience."

Neil Chue Hong, founding director of the Software Sustainability Institute at the University of Edinburgh, UK, helped to establish the FAIR4RS principles. Chue Hong says that, over the past few decades, increased reliance on software is among the largest changes to data science , such that almost every branch of research now uses software in some way. As a result, the institute advocates for the fundamental importance of providing scientists with training on best practices when using research software. "It's now very hard to analyse or visualize data without software, and at the same time, it's very hard for software to exist without high-quality data," he says.

Just as data should come with a metadata or README file that contains information about the data set itself, software and algorithms should also have good documentation, including which version a person used. That's especially true for artificial-intelligence research. For instance, HuggingFace, a model-sharing service based in New York City, encourages researchers to create 'model cards' that provide key information about AI tools, including their intended use, performance metrics, training data and limitations.

The decade to come

Although widely accepted today, FAIR had its teething troubles. Early on, some researchers treated the principles as a box-ticking exercise. In 2017, the development team released clarifying remarks to address the issue[7]. The authors noted that FAIR does not require data to be entirely open – instead, accessiblity is the goal. They also clarified that although the team members had life-sciences backgrounds, the guidelines could be adopted to fit any discipline.

"More people in the scientific community appreciate that data can be available yet still have some guard rails."

The experience helped Mons to land on a new vision for FAIR's future, rooted in the premise that the FAIRest data are those making a demonstrable difference. Mons is now thinking about how FAIR can best be applied, "such that we're not just focusing on getting the data uploaded, but making insights with them", he says.

Mons is a founding member of a project launched in 2024 called the Leiden Initiative for FAIR and Equitable Science (LIFES), based on the idea that researchers should be able to access analytical tools remotely, rather than having to get a copy themselves. Such 'data visiting' keeps information accessible but gives its creators more control over how and when sensitive data can be used. This layer of security can be especially important when dealing with health data or data from marginalized communities, and many aspects of data visiting dovetail with another set of guidelines, the CARE principles[8]. These were developed by Indigenous researchers to enhance data sovereignty, which gives Indigenous Nations the right to govern the collection, ownership and application of their data.

"There was initially some conflict between FAIR and CARE when it came to defining what open data really are," says Maui Hudson, director of the Māori-led Te Kotahi Research Institute in Hamilton, New Zealand, and a developer of CARE. "It's good to see more people in the scientific community appreciate that data can be available yet still have some guard rails."

In the past few years, Mons has, at times, redefined FAIR to stand for 'Fully AI-Ready', recognizing how AI has become ubiquitous in scientific workflows. Other researchers, including Huerta, have adapted the FAIR principles to align with AI-based systems, too. Current FAIR AI guidelines include sections on building data frameworks that can more easily interface with AI, such as by identifying the data types most used by machine-learning algorithms and creating application programming interfaces that allow AI-driven applications to communicate and share data with each other[2].

Julie Jurgens, the scientific communication and curation manager of the Knowledge Portal Network from the Broad Institute of MIT and Harvard in Cambridge, Massachusetts, is similarly looking ahead to how FAIR can be adapted to an AI-driven world. Jurgens is part of an initiative called the Common Fund Data Ecosystem (CFDE), which is led by the US National Institutes of Health (NIH). It seeks to leverage data generated through research backed by the Common Fund – an NIH funding programme that supports short-term projects involving more than one of the agency's institutes and centres – to accelerate discovery. Building the federated ecosystem has meant reconciling differences in how people typically think about data and keeping things FAIR.

Towards this end, the CFDE has created a metadata model that helps members to describe files, samples, subjects, assays and experiments in an interoperable, machine-readable way. The CFDE also offers shared submission pipelines, validation workflows, FAIRness assessment tools and high-performance computing environments for large-scale analytics.

Reflecting on ten years of FAIR, Mons says that he is proud of how the academic community has embraced and adapted the idea. "It keeps me optimistic," he says, "about the future of open and FAIR data – and also motivates me to continue my work strengthening our systems to meet the challenges of the future."

Amanda Heidt is a freelance journalist in southeast Utah.

1. Wilkinson, M. *et al. Sci. Data* **3**, 160018 (2016).
2. Huerta, E. A. *et al. Sci. Data* **10**, 487 (2023).
3. Belliard, F. *et al. PLoS Comput. Biol.* **19**, e1011668 (2023).
4. Chen, Y. *et al. Sci. Data* **9**, 31 (2022).
5. Sonabend, R., Gruson, H., Wolansky, L., Kiragga, A. & Katz, D. S. *PLoS Comput. Biol.* **20**, e1012045 (2024).
6. Barker, M. *et al. Sci. Data* **9**, 622 (2022).
7. Mons, B. *et al. Inf. Serv. Use* **37**, 49–56 (2017).
8. Carroll, S. G. *et al. Data Sci. J.* **19**, 43 (2020).