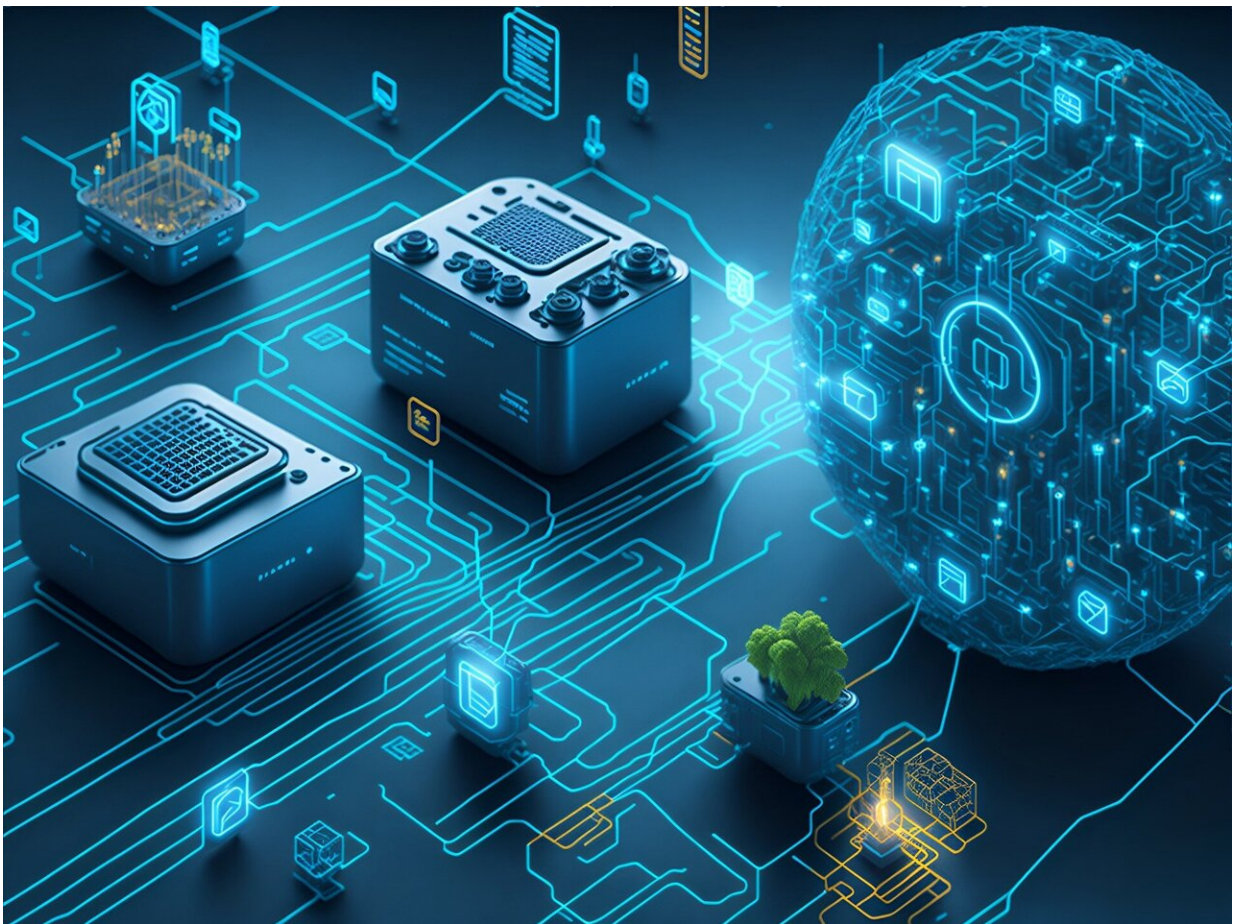# AI and new standards promise to make scientific data more useful by making it reusable and accessible

August 22 2023, by Bradley Wade Bishop

Credit: Pixabay/CC0 Public Domain

Every time a scientist runs an experiment, or a social scientist does a survey, or a humanities scholar analyzes a text, they generate data. Science runs on data—without it, we wouldn't have the James Webb Space Telescope's stunning images, disease-preventing vaccines or an evolutionary tree that traces the lineages of all life.

This scholarship generates an unimaginable amount of data—so how do researchers keep track of it? And how do they make sure that it's accessible for use by both humans and machines?

To improve and advance science, scientists need to be able to reproduce others' data or combine data from multiple sources to learn something new.

Any kind of sharing requires management. If your neighbor needs to borrow a tool or an ingredient, you have to know whether you have it and where you keep it. Research data might be on a graduate student's laptop, buried in a professor's USB collection or saved more permanently within an online data repository.

I'm an information scientist who studies other scientists. More precisely, I study how scientists think about research data and the ways that they interact with their own data and data from others. I also teach students how to manage their own or others' data in ways that advance knowledge.

Research data management

Research data management is an area of scholarship that focuses on data discovery and reuse. As a field, it encompasses research data services, resources and cyberinfrastructure. For example, one type of infrastructure, the data repository, gives researchers a place to deposit their data for long-term storage so that others can find it. In short,

research data management encompasses the data's life cycle from cradle to grave to reincarnation in the next study.

Proper research data management also allows scientists to use the data already out there rather than recollecting data that already exists, which saves time and resources.

With [increasing science politicization](#), many national and international science organizations have upped their [standards for accountability and transparency](#). [Federal agencies](#) and other major research funders like the [National Institutes of Health](#) now prioritize research data management and require researchers to have a data management plan before they can receive any funds.

Scientists and data managers can work together to redesign the systems scientists use to make data discovery and preservation easier. In particular, [integrating AI](#) can make this data more accessible and reusable.

Artificially intelligent data management

Many of these new standards for research data management also stem from an increased use of AI, including [machine learning](#), across [data-driven fields](#). AI makes it highly desirable for any data to be machine-actionable—that is, usable by machines without human intervention. Now, scholars can consider machines not only as tools but also as potential autonomous data reusers and collaborators.

The key to machine-actionable data is metadata. [Metadata](#) are the descriptions scientists set for their data and may include elements such as creator, date, coverage and subject. Minimal metadata is minimally useful, but correct and complete standardized metadata makes data more useful for both people and machines.

It takes a cadre of research data managers and librarians to make machine-actionable data a reality. These information professionals work to facilitate communication between scientists and systems by ensuring the quality, completeness and consistency of shared data.

The FAIR data principles, created by a group of researchers called FORCE11 in 2016 and used across the world, provide guidance on how to enable data reuse by machines and humans. FAIR data is findable, accessible, interoperable and reusable—meaning it has robust and complete metadata.

In the past, I've studied how scientists discover and reuse data. I found that scientists tend to use mental shortcuts when they're looking for data—for example, they may go back to familiar and trusted sources or search for certain key terms they've used before. Ideally, my team could build this decision-making process of experts and remove as many biases as possible to improve AI. The automation of these mental shortcuts should reduce the time-consuming chore of locating the right data.

Data management plans

But there's still one piece of research data management that AI can't take over. Data management plans describe the what, where, when, why and who of managing research data. Scientists fill them out, and they outline the roles and activities for managing research data during and long after research ends. They answer questions like, "Who is responsible for long-term preservation," "Where will the data live," "How do I keep my data secure," and "Who pays for all of that?"

Grant proposals for nearly all funding agencies across countries now require data management plans. These plans signal to scientists that their data is valuable and important enough to the community to share. Also, the plans help funding agencies keep tabs on the research and investigate

[any potential misconduct](#). But most importantly, they help scientists make sure their data stays accessible for many years.

Making all research data as FAIR and open as possible will improve the scientific process. And having access to more data opens up the possibility for more informed discussions on [how to promote](#) [economic development](#), improve the stewardship of natural resources, enhance public health, and how to responsibly and ethically develop technologies that will improve lives. All intelligence, artificial or otherwise, will benefit from better organization, access and use of research data.

This article is republished from [The Conversation](#) under a Creative Commons license. Read the [original article](#).

Provided by The Conversation