Science Societies

Protecting Privacy While Making Data Open in Agricultural Research

By Kaine Korzekwa

March 1, 2023

USDA Photo by Lance Cheung.

The open science movement is a multi-faceted set of pillars that works to make the process of scientific research more transparent, accessible, efficient, reliable, and responsive to societal challenges. The pillars encompass open access, open data, open science evaluation, open science policy, open science tools, and more.

The idea is bold, potentially revolutionizing the way research is conducted and communicated. As the practices of open science—particularly open data requirements—continue to ripple through all areas of the scientific workforce, some scientists in specific areas are finding unique challenges that they must navigate to meet open science standards and requirements. Tackling these growing pains will be essential to the continued advancement of the open science movement.

“Open science, also being called open research, is this ongoing transition in how research is being performed and how the knowledge from that research is being shared,” explains Kathleen Yeater, a statistician with USDA-ARS who is currently serving as ASA’s editor-in-chief. “Fast-paced advances in digital technology and the knowledge of how much open science can benefit the greater good has motivated many stakeholders, such as funders, to require researchers to adhere to certain open data standards.”

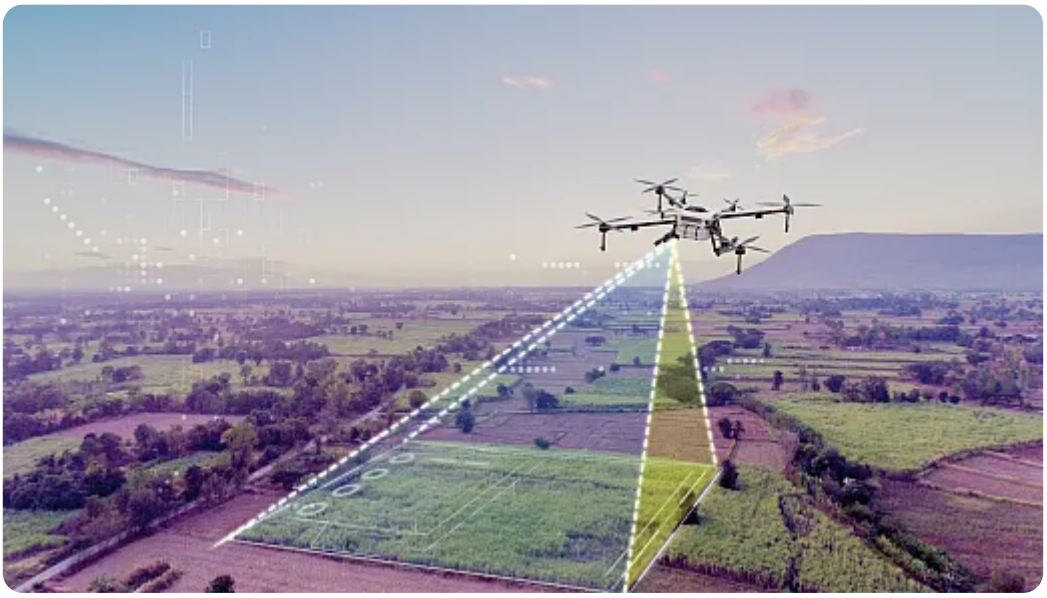
The history of open science can be traced all the way back to the conception of scientific journals, which allow for the publishing and sharing of scientific findings across the globe. More modern definitions, such as that of open data, refer to requirements set by funding agencies, as well as journals and researchers, to share research data in a way that is free and open to all interested parties.

The National Center for Biotechnology Information began to manage genomic data repositories, such as the Gene Expression Omnibus, as early as 2002. In 2013, during the Obama Administration, the Office of Science and Technology Policy directed federal agencies that receive more than $100 million in research and development funds to develop plans to make the results of federally funded research open and available to the public. From there, major journal publishers began to adopt various data availability mandates to ensure the validity and quality of the research being published.

To maintain consistency in open data, a set of standards called the FAIR principles was conceived. The principles add an expectation that data be “findable, accessible, interoperable, and reusable” to ensure that the ethos behind open data bears out in reality. The benefits of open data are immense, and some reports suggest that strategies like data archiving could generate a 250-fold increase in scientific publications compared with federal funding alone. Research data that are easily searchable and useable can save time and allow researchers to better build on each other’s work. They also allow for quality checks to ensure findings are reproducible and verifiable.

To Yeater, this opens up incredible possibilities in the publishing of data. She imagines an open data infrastructure where data are publishable and citable with a digital object identifier (DOI), so researchers receive credit for their work. “From my perspective as a researcher, being able to have my high quality open data cited would be as important and valuable for me and my career as an actual published article,” she says. “It would allow us to begin to treat data as a first-class research product.”

As with any large movement, there are unique issues to navigate. For instance, although data may be in a state where they can be found and theoretically used, that does not guarantee they are useful. Due to a lack of resources, awareness, and support, many scientists post data that fulfill funding entity requirements but are not actually usable by others.



Experience and research show that farmers have valid privacy concerns about data from their farms being used and shared inappropriately. Photo courtesy of Adobe Stock/angkhan.

Fair to Science and Fair to Farmer Privacy

In a recent Agricultural and Environmental Letters commentary (https://doi.org/10.1002/ael2.20062), Joby Czarnecki, associate research professor, and Mary Ann Jones, associate professor—both at Mississippi State University—point out several issues facing researchers working with geospatial data for on-farm research and offer potential solutions that can be explored. “The key point is that we want open data and think it is an incredibly important tool,” Czarnecki says. “In most

instances, it makes great sense. It is a bit harder when you start talking about personal data, like farm data. We don't really have a good guide for what I am supposed to do with an individual landowner's farm data that's fair to science but also fair to the farmer and their privacy." She says blanket language about open data requirements from funders, journals, and others causes problems for data that may contain private information that is difficult to de-identify. Czarnecki draws a contrast with fields in medicine where the human health stakes may be high. Over time and by necessity, these fields have made information de-identifiable and shareable. The same evolution is just beginning for agricultural data, she says.

"There are many solutions out there that work for lots of fields of research, but something many don't appreciate is how much harder agricultural data can be," Czarnecki explains. "We can't de-identify it like medical data, and an algorithm to disguise the data has the potential to be undone.

We have these odd, unique challenges."

Researchers feel the need to honor the request of a collaborating farmer in order to foster a good relationship and continue their work. Photo by Aaron Hird, USDA-NRCS.

Experience and research show that farmers have valid privacy concerns about data from their farms being used and shared inappropriately. Czarnecki shares examples of farmers asking for their house to be cropped out of drone imagery. Researchers feel the need to honor the request of a collaborating farmer in order to foster a good relationship and continue their work. Further issues arise when imagery data become combined with management practices and income. While data in a single dataset can be de-identified, there are several reports of multiple datasets being combined later on and the data being re-identified. "In theory, someone could access information at a state office and learn who owns a property and access all kinds of characteristics about the property and the owner," Czarnecki says. "Farmers have concerns, and some reporting has shown, their information could be used to sell them products or harass them for certain agricultural practices that are not fully understood by those outside of agriculture."

Three Potential Solutions

Czarnecki and Jones highlight three potential solutions from the literature and discuss their pros and cons. All three center on a researcher not needing to release all of their data, but instead share
enough to show whether the data are valid and may be helpful to other researchers. The solutions include providing metadata, establishing a data enclave, and requiring a data subset. Having researchers provide open metadata but controlling access to the data is one option. It would provide enough documentation about a dataset to help

others discern its quality and credibility and
determine if they want to contact the researcher to learn more. “Metadata allows someone to understand what data were collected, how, and why to determine if they would be able to use it,” Czarnecki says. “Issues arise when researchers don’t actually respond to requests for shared data and some scientists have difficulty providing high quality metadata, especially in fields of research that produce huge amounts of data.”

The idea of a controlled-access data enclave would allow researchers to access and use data but not directly edit and keep them—a kind of “read only” format. While promising, Czarnecki and Jones note how past trials of this approach quickly ran into accessibility issues as well as a lack of resources and funding for data storage and management. Czarnecki says that, to her, the most promising proposition out there is the concept of providing a curated data set, rather than a whole data set. This meets two goals: researchers have to prepare and annotate a smaller amount of data but still provide enough data to give potential users an idea of whether the full data set will be useful. “If the goal of the journal or funder is to validate that my data are sound, that it can provide the outcome I say it can, and that the method is repeatable, I think that is easily accomplished with the curated data set approach,” she says. “If it piques a researcher’s interest, they can contact me for more data. It allows us to focus on quality over quantity in terms of preparing and annotating our data.”

The approach also allows for reduced risk and privacy concerns. For example, if Czarnecki is working with four landowners, and three of them are uncomfortable having their data shared widely, she can just share the data from one, providing flexibility.

Who Decides Which Methods Are Adopted?

Who decides which method or methods are adopted? Czarnecki and Yeater both say that the research community has a strong grassroots voice, especially because it is researchers themselves who volunteer to be editors of many journals. It will also depend on the needs of specific fields of research, some of which have unique challenges.

For example, Yeater says, plant breeders who are focused on releasing information about new cultivars often run into issues with protected or private information that can conflict with open data requirements.

“I believe in the community’s ability to be able to determine what is right for them, and I want to drive support for our members,” Yeater says. “We want to be on the side of recognizing and rewarding open practices. This is going to be a long, very thoughtful process that involves community engagement and a culture shift. I view it as an opportunity to be really proactive and value conversations and commentaries that explore these issues.”

What Is the Current Data Policy for ASA, CSSA, and SSSA Publications?

Our journal data policy is found in the ASA, CSSA, and SSSA Editorial Policies document (agronomy.org/publications/journals/author□resources/editorial□ policies). In a nutshell, we encourage the “storage and availability of data necessary to understand and evaluate phenomena reported in our publications.”

All ASA, CSSA, and SSSA journals also encourage authors to include a data availability statement. Data papers and all papers in *The Plant Genome* and *Urban*

Agriculture & Regional Food Systems require such a statement.

In addition, we urge authors to store or archive their data in domain repositories that are widely recognized and available to the community.

Many repositories assign DOIs, which are “persistent” URLs (they never change). This means datasets deposited there are easier for people to find. Generalist repositories accept datasets from a number of disciplines. One such generalist repository is Dryad, which our journal authors can submit to as part of our submission process at no cost.

Other generalist repositories that assign DOIs include figshare and Zenodo. Some institutional repositories also can assign DOIs or other persistent identifiers such as Handle; you can ask your institution if it does this. And there is an ever-growing number of discipline-specific (specialist) repositories, such as PANGAEA, which accepts data from the fields of earth, environmental, and life sciences.

Different journals have different policies, so be sure to check the individual journal’s author instructions for details (agronomy.org/publications/journals/author-resources). For example, data papers published in our journals require that the data must be publicly and freely available.

In this shifting landscape of open science, we encourage authors to make their data freely available when possible, taking into account any confidentiality concerns.

Resources: agronomy.org/publications/journals/author-resources/editorial-policies; agronomy.org/publications/journals/author-resources

DIG DEEPER

View the original commentary, “The Problem with Open Geospatial Data for On-Farm Research,” in *Agricultural & Environmental Letters* at https://doi.org/10.1002/ael2.20062
