World view

By Jessica Farrell

To make data open, stop overlooking librarians

Digital archivists are experts at tackling the challenges of making data accessible and open. We can help to smooth the transition.

The 'Year of Open Science', as declared by the US Office of Science and Technology Policy (OSTP), is now wrapping up. This followed an August 2022 memo from OSTP acting director Alondra Nelson, which mandated that data and peer-reviewed publications from federally funded research should be made freely accessible by the end of 2025. Federal agencies are required to publish full plans for the switch by the end of 2024.

But the specifics of how data will be preserved and made publicly available are far from being nailed down. I worked in archives for ten years and now facilitate two digital-archiving communities, the Software Preservation Network and BitCurator Consortium, at Educopia in Atlanta, Georgia. The expertise of people such as myself is often overlooked. More open-science projects need to integrate digital archivists and librarians, to capitalize on the tools and approaches that we have already created to make knowledge accessible and open to the public.

Making data open and 'FAIR' — findable, accessible, interoperable and reusable — poses technical, legal, organizational and financial questions. How can organizations best coordinate to ensure universal access to disparate data? Who will do that work? How can we ensure that the data remain open long after grant funding runs dry?

Many archivists agree that technical questions are the most solvable, given enough funding to cover the labour involved. But they are nonetheless complex. Ideally, any open research should be testable for reproducibility, but re-running scripts or procedures might not be possible unless all of the required coding libraries and environments used to analyse the data have also been preserved. Besides the contents of spreadsheets and databases, scientific-research data can include 2D or 3D images, audio, video, websites and other digital media, all in a variety of formats. Some of these might be accessible only with proprietary or outdated software.

Librarians have many tools that can help, such as ReproZip, created by Rémi Rampin and supported by Vicky Rampin at New York University in 2013. This software brings together into one package all the data files, libraries, environmental variables and options needed to reproduce research. The open-source software BitCurator has supported digital archiving work since 2011. Thanks to years of work by many archivists, the US Library of Congress and the UK National Archives both maintain registries of file formats and what software is needed to open them.

Legal and organizational barriers are trickier. For example, in the United States, under the 1998 Digital Millennium Copyright Act, a library couldn't break a digital lock on software, even for preservation or research. A long-lost password, a defunct authentication server or a broken dongle could render data inaccessible. Thanks to advocacy by the Software Preservation Network, updated rules allow libraries to break those locks to preserve software in their collections, ensuring long-term access to data. The Software Preservation Network continues to press for policy changes that enable the preservation of and access to software.

> "More open-science projects need to integrate librarians and digital archivists, to capitalize on tools we have already created."

There is also no one body to provide oversight for ensuring data are open. Funders should consider how they could support the formation of organizations that do this, made up of both scientists and information scientists, to help to coordinate across projects and avoid duplications.

All of this requires people to overcome outdated misconceptions of librarianship. If you're a scientist who has never thought about archivists before, there might be cultural reasons for that. Information science is a feminized field, and archivists are often underpaid and perceived as administrative support staff, not co-creators in the knowledge-production process. Archives are often imagined as boxes of dusty papers, but most archives today maintain vast amounts of digital data. Information management is an academic discipline and should be treated as such.

Fortunately, there are examples of fruitful partnerships between researchers and archivists. NASA's Year of Open Science and the Scientific Information Service at CERN near Geneva, Switzerland, co-hosted an open-science summit in July. My colleague Paul Gignac, a vertebrate palaeontologist at the University of Arizona in Tucson, sought out the expertise of digital archivists when setting up the NSF-funded Non-Clinical Tomography Users Research Network. The project is investigating how to preserve 3D-imaging data sets and how to track important contextual information, such as where the data came from and notes on reproducibility. Gignac found that using information-science tools and standards — such as including metadata about how materials were preserved — helped to ensure that data were FAIR without reinventing the wheel. He also collaborates with the Data Curation Network, a community hub hosted by the University of Minnesota in Minneapolis, which anyone can join.

Many digital archivists and scientists share a vision of a world in which reliable open data are maintained, quality scientific information is accessible regardless of income or location and — as has recently become important — large language models can be trained on well-curated open data instead of on data of unverified quality used without permission. The expertise of digital archivists can help scientists and society to extract maximum benefit from the transition to open access.

Jessica Farrell is a community facilitator at the Educopia Institute in Atlanta, Georgia.
e-mail: jess.aileen.farrell@gmail.com

The author declares competing interests; see go.nature.com/4adyusc.

JESSICA FARRELL