THE LIBRARY REBOOT

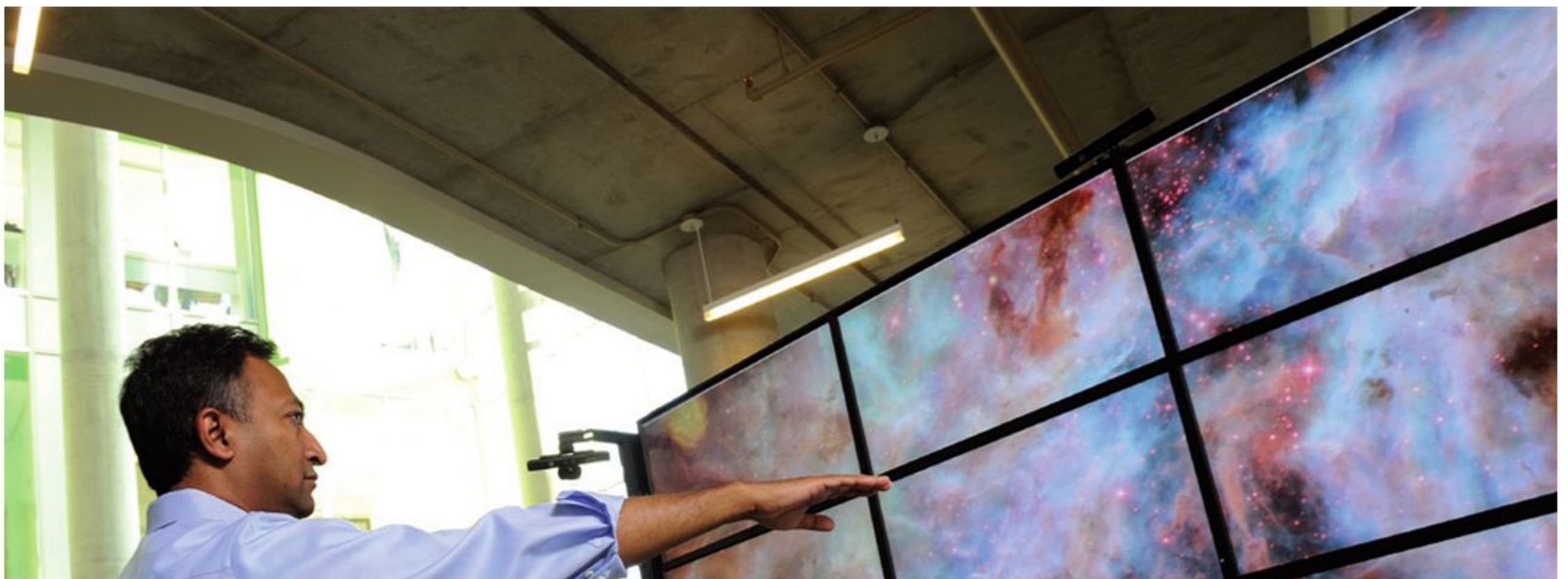
As scientific publishing moves to embrace open data, libraries and researchers are trying to keep up.

BY RICHARD MONASTERSKY

A few passing students do a double take as Sayeed Choudhury waves his outstretched right arm. In his crisply pressed dress shirt and trousers, the engineer looks as if he is practising dance moves in slow motion. But he is really playing with astronomical data.

Standing in a US$32-million library building opened last year at Johns Hopkins University in Baltimore, Maryland, Choudhury faces a 2-metre-by-4-metre 'visualization wall' of television screens. Pointing with his arm, he selects a picture of the Ring Nebula out of 40 images from the Hubble Space Telescope. Choudhury spreads his hands in a welcoming gesture and the nebula's rim of glowing orange gas fills the frame.

This wall is the brainchild of computer scientist Greg Hager and Choudhury, who directs digital research and curation at the library. For $30,000, they and their team patched together monitors, processors and the Microsoft Kinect system that recognizes arm and body gestures. They placed the wall in the library last October as an experiment, allowing students and researchers to explore a few of the university's data sets, from star



WILL KIRK/JHU HOMEWOOD PHOTOGRAPHY

Sayeed Choudhury demonstrates the visualization wall, part of Johns Hopkins University's drive to transform how its libraries and researchers deal with data.

© 2013 Macmillan Publishers Limited. All rights reserved

systems to illustrated medieval manuscripts.

"As we create more and more digital content, there's a question of how do you get people to even realize we have it and then interact with it in new ways," says Choudhury, who thinks that the wall is starting to catch on. One chemical engineer wants to use it to visualize and manipulate molecules, and astronomers hope that it could help to train students in categorizing galaxies. By providing alternative ways to explore and share data, says Choudhury, the wall "is a new form of publishing".

Around the world, university libraries are racing to reinvent themselves to keep up with rapid transformations in twenty-first-century scholarship. They still do a brisk business in purchasing books, licensing access to academic journals and providing study spaces and research training for students. And libraries are increasingly helping teachers to develop courses and adopt new technologies. But for working scientists, who can now browse scientific literature online without leaving their desks, much of this activity goes unseen. For many, libraries seem to be relics that no longer serve their needs.

"I SEE US MOVING UP THE FOOD CHAIN AND BEING CO-CONTRIBUTORS TO THE CREATION OF NEW KNOWLEDGE."

THE NEW DATA WRANGLERS

That could soon change. At Johns Hopkins and many other top universities, libraries are aiming to become more active partners in the research enterprise — altering the way scientists conduct and publish their work. Libraries are looking to assist with all stages of research, by offering guidance and tools for collecting, exploring, visualizing, labelling and sharing data. "I see us moving up the food chain and being co-contributors to the creation of new knowledge," says Sarah Thomas, the head of libraries at the University of Oxford, UK.

It is not yet clear how successful this reinvention will be, especially given the tight budgets facing both libraries and researchers. And as they step into the data-curation business, libraries are entering a crowded market of commercial publishers, information-storage companies and discipline-specific data repositories such as GenBank, which archives DNA sequences. But many say that libraries have a natural role in the data world, and that their importance will only grow with the push to unlock the products of research.

Last month, US President Barack Obama's administration ordered granting agencies to ensure that the public can access publications and data generated by federally funded research. "This is going to have significant repercussions and result in much greater appreciation and support for the need to preserve data and make it available for scientific use," says William Michener, an information scientist at the University of New Mexico libraries in Albuquerque. "Libraries are really critical stakeholders in this new landscape. They're the first line of defence when faculty members have a problem with managing their data."

The research team at the Hopkins' Sheridan Libraries is one of the leaders in planning for this transformation, thanks in part to more than a decade of experience managing data from the Sloan Digital Sky Survey, which has mapped nearly one million galaxies. Choudhury is also principal investigator on a $9.4-million grant from the US National Science Foundation (NSF) to develop the Data Conservancy, a multi-institution programme that is researching and building tools for data curation. That grant enabled Johns Hopkins to launch a fee-based service in 2011 to help researchers manage their data.

University scientists were not thrilled with the idea at first. During a 2011 meeting to describe the efforts, some rebelled against what seemed to be a mandatory data-management charge — "sort of like a tax on proposals", says Noah Cowan, a mechanical engineer at Johns Hopkins. It did not help that Cowan and his colleagues were already dealing with new data-management obligations: earlier that year, the NSF had started requiring that grant-seeking researchers make clear how they would disseminate and share their data. Choudhury had to work hard to explain that Johns Hopkins' data service would be voluntary and useful.

"Preservation is not a big selling point for researchers," says Betsy Gunia, a data-management consultant at the university. "Some of it is lack of awareness."

Cowan, who studies animal biomechanics, eventually stepped forward as one of the service's first customers, thinking that he could improve on his usual practices. In his office, he pulls up an example of the kind of data he generates: high-speed videos of a knifefish swimming in a simulated current. His team records the fish's fin movements and his neuroscientist colleagues measure nerve signals to study how the animal controls its position in the water.

Although the science is cutting edge, Cowan's approach to data is relatively old-school. When a study is complete, he stores the video and analyses on a hard drive on his shelf. And like many other researchers, he shares his data on a case-by-case basis in response to requests. Those methods have generally sufficed, but when a graduate student wanted to reanalyse some 7-year-old studies last summer, it took a few months to make sense of the data. They were kept separately from the analysis code and there were multiple versions of code to sort through. The poor quality of the metadata — information that describes the data — "made it a treasure hunt", says Cowan.

So when he started drafting a new NSF proposal to explore the neural activity of singing birds, he sat down with Gunia and another information specialist. Together, they developed a plan to organize the project's data and make some available to outside researchers. If the NSF funds the project, the Johns Hopkins service will curate and store Cowan's data for five years under a renewable contract.

The curation process is much more involved than simply storing data online through a service such as Dropbox. Cowan will supply 'readme' files of metadata as well as any scripts used to collect or process the data. The service will also help him to label the data with a unique and permanent reference, such as the digital object identifier (DOI) numbers used by publishers — a vast improvement over web links to a data set, which can break, resulting in the all-too-familiar '404 error' message. And with a persistent identifier, the data become directly citable by others.

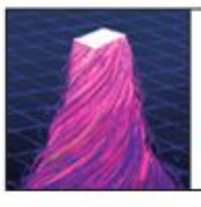
The Johns Hopkins data-management team also agrees to safeguard against problems such as the storage media degrading, files getting corrupted and data formats becoming obsolete — protections not offered by many existing university repositories, which primarily house digital documents. "I don't have to suffer the problem of bit rot," says Cowan.

If the grant is funded, the services will amount to roughly 2% of direct costs, but Cowan doesn't mind paying. "My time is better spent mentoring students and collecting and analysing data than running my own long-term data archive," he says.

CONCERN OVER CURATION

Many scientists are too busy or lack the knowledge to tackle data-management on their own. In a survey of 1,300 scientists completed in 2010, more than 80% said that they would use other researchers' data sets if they were easy to reach, but only 36% said that others could access their data easily (see 'The data gap') (C. Tenopir *et al. PLoS ONE* **6,** e21101; 2011).

Scientists who do their own data curation can take advantage of many new options, such as DataONE, an international network for preserving and sharing data that was developed through a $20-million grant from the NSF, which Michener is leading. Another non-profit option is Dryad,


THE FUTURE OF PUBLISHING
A *Nature* special issue.
nature.com/scipublishing

© 2013 Macmillan Publishers Limited. All rights reserved

THE DATA GAP

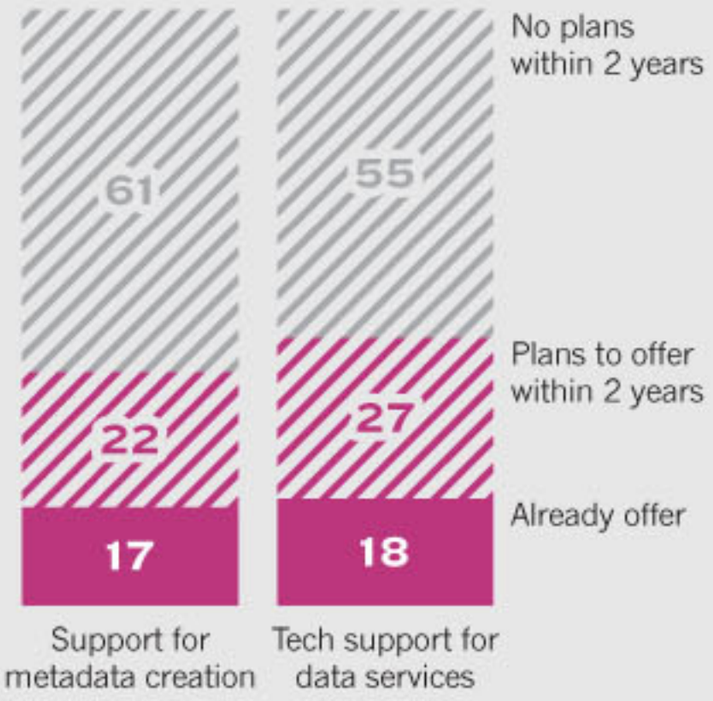
A survey of more than 1,300 US scientists in 2010 showed an appetite for sharing data but significant hurdles that kept many from doing it.

67%
Said lack of access to others' data is a major impediment to scientific progress.

39%
Said their organization has an established process for long-term data storage.

22%
Said their organization or project provides funds for long-term data management.

A survey of US research universities showed that many libraries are gearing up to provide data services.



SOURCE: LEFT: C. TENOPIR *ET AL. PLOS ONE* **6**, E21101 (2011); RIGHT: TENOPIR/ALLARD/SANDUSKY/BIRCH/NSF DATAONE PROJECT

a repository that helps researchers to identify data sets, store them and connect them with publications. Data services are also available from a range of companies, including figshare in London (owned by Macmillan Publishers, the parent company of *Nature*) and the Data Citation Index, launched last year by Thomson Reuters in New York.

However, libraries are particularly well positioned to fill the data-management gap, says Michener. They have long experience helping faculty members and they are not likely to disappear, he says. "Libraries have a significant amount of trust capital."

And many are interested in entering the data-management game. Carol Tenopir, an information scientist at the University of Tennessee in Knoxville who ran the 2010 survey of scientists' data habits, also conducted an as-yet unpublished study of libraries at more than 100 US research universities in 2011–12. Although fewer than 20% offered data services at the time, nearly 40% had plans to help scientists to curate and store their data within two years, says Tenopir.

Oxford's Bodleian Libraries are already developing a fee-based approach with different tiers of storage, from completely closed levels for sensitive information such as patient data to more open tiers for publicly accessible data and metadata. A survey of Oxford researchers last year indicated that they collectively would have about 3 petabytes (3 million gigabytes) of data to deposit in the repository's first year — roughly double what is currently stored in Oxford's central file system. But they will probably not deposit the whole amount, says Wolfram Horstmann, who is developing the digital services for the Bodleian. The survey did not mention that researchers would have to document their data and pay for storage, at a rate of about £5,000 (US$7,500) per terabyte (1,000 gigabytes).

Other universities are exploring different approaches for curating research data. Stanford University in California, for example, is piloting a data-management service and a repository in which researchers can deposit their own data, with no cost for small and simple items. Yet many universities lack the resources to house their own repositories and instead will help researchers to find appropriate existing ones. "It's just the very top tier of research institutions that are going to be able to be data repositories," says Tenopir.

The push to share data is accelerating in several countries. The Australian government has invested AU$75.5 million (US$78.3 million) to establish the Australian National Data Service, run from Monash University in Melbourne. Monash librarian Cathrine Harboe-Ree says that the service is helping Australian universities to identify and publish all kinds of information — "everything from the petabytes of data from a synchrotron to the quite modest material from an oral research project".

The data-curation efforts at Monash, Johns Hopkins and other libraries dovetail with what many people say is a revolution in scientific publishing: the move away from narrative, text-based papers as the major output of research. "The classical scholarly publishing model has reached its limit because it mostly hasn't changed in the past 300 years," says Jan Brase, an information researcher at the German National Library of Science and Technology in Hanover and managing director of DataCite, an international organization focused on describing and citing data sets.

In the future, scientific output could be measured using all types of data, from spreadsheets of observations to algorithms and analysis tools. "Until recently, data have been considered a second-class citizen in the science and publishing world, and that's all about to change," says Michener.

FROM PAPERS TO PRODUCTS

A step in that direction came this year, when the NSF altered its proposal guidelines by allowing researchers to list 'products' that they have created, such as data sets and software, instead of just publications. At a congressional hearing this month, Choudhury and others made the case that public access to scientific data is a matter of national competitiveness. And proponents of open data argue that it will help to expose fraud and mistakes in research.

In negotiating all these changes, libraries face huge challenges, including shrinking budgets and the rising cost of journal subscriptions. "The combination of the paradigm shift and fiscal pressure has led some to use the word crisis," says Brian Schottlaender, chief librarian at the University of California, San Diego, whose budget shrank by 21% between 2008 and 2012. Schottlaender himself takes a less dire view, saying that libraries are at a "crossroads".

It is too early to say whether libraries will succeed at redefining themselves in the digital age, says Eefke Smit, director of standards and technology at the International Association of Scientific, Technical & Medical Publishers in Amsterdam, the Netherlands. But "it definitely seems that some are successful in setting up new things".

If the emerging complex data environment is any predictor, libraries will one day be part of a vast ecosystem dealing with information curation. But everybody hopes that the borders between repositories will be seamless, so that a researcher running a query from his or her desk can pull up data from around the world.

Many portray the new focus on data as a big change for libraries, but Thomas says that it is not a huge departure from what they have been doing for centuries: organizing information, preserving it and making it available to scholars. Scientific data sets are more complicated, she says, "but in some respects they're no different from a page in a medieval manuscript". ■

Richard Monastersky *is a features editor for* Nature *in Washington DC.*

© 2013 Macmillan Publishers Limited. All rights reserved