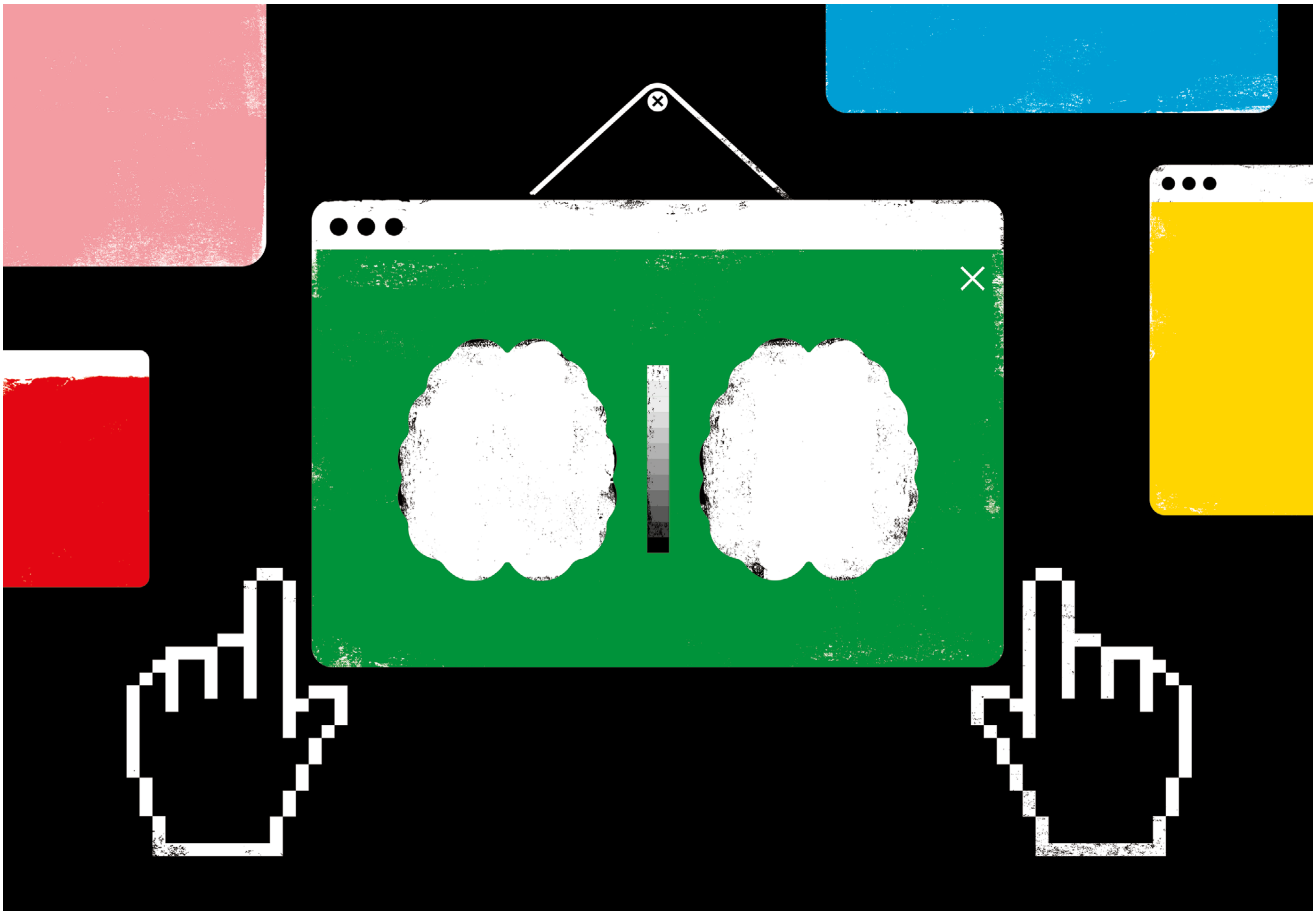

ILLUSTRATION BY THE PROJECT TWINS

A HOME FOR EVERY IMAGING DATA SET

Repositories let researchers store, share and access life-science images – and maybe even extract new findings. **By Amber Dance**

When Sjors Scheres set out to develop a tool to reverse flaws in cryo-electron microscopy images, he needed lots of data on which to test it. So Scheres, a structural biologist at the MRC Laboratory of Molecular Biology (LMB) in Cambridge, UK, turned to the Electron Microscopy Public Image Archive (EMPIAR), a database of raw images. There he downloaded, for free, data collected by the lab of Gabriel Lander, a structural biologist at Scripps Research in La Jolla, California.

Using his new technique, Scheres was able to squeeze sharper images from those data, improving[1] the resolution of one structure from 3.1 ångströms to 2.3 ångströms.

"That's precisely why we posted the data," says Lander. "We knew some brilliant people out there would be able to improve on our processing."

Services such as EMPIAR give researchers a central location in which to store, share and access a rapidly expanding corpus of biological images. "The data aren't just one picture any more," says Joshua Vogelstein, a neurostatistician at Johns Hopkins University in Baltimore, Maryland. Movies, 3D images and microscope-based screening data can take up gigabytes or terabytes of storage, and can't be e-mailed back and forth in the same way as individual TIFF or JPEG files. Moreover, grant agencies and journals increasingly require scientists to make their data available to all, but don't necessarily offer to host them. EMPIAR and its kin fill that gap, and often provide a digital object identifier or other citation so researchers can get credit for their data.

"Are you struggling to load your images?" asks Forrest Collman, a neuroscientist at the Allen Institute for Brain Science in Seattle, Washington. "Are you particularly struggling to share?" If so, he says, "looking into this kind of service makes sense for you".

In 2019, when Collman spotted an odd-looking neuron in one of his electron-microscopy data sets[2], it was easy for him to send a colleague a link to that spot in the data repository, rather than a bulky file. She noticed another unique feature, and Collman identified a few similar cells. They might turn out to be a new type of neuron, Collman says.

There are a number of other image warehouses available, among them the Image Data Resource (IDR). Both it and EMPIAR are hosted by the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) in Hinxton, UK. Further options include, but are not limited to, NeuroData, a platform that Vogelstein set up to host neuroanatomy files, and the Systems Science of Biological Dynamics (SSBD) database at Japan's RIKEN network of research institutes. Advocates expect these platforms to follow the model of established DNA- and protein-sequence resources such as GenBank and the Protein Data Bank, which have powered an array of analyses and spawned the field of bioinformatics.

"We're very early days," says Jason Swedlow, a quantitative cell biologist at the University of Dundee, UK. But he expects big benefits, both for scientists who download large image sets to feed data-hungry machine-learning algorithms and for those who might make new discoveries in others' data.

Share and share alike

It was a data-hungry scientific community that drove Kate McDole, a developmental biologist at the LMB, to use an image database.

McDole, then working at the Howard Hughes Medical Institute's Janelia Research Campus in Ashburn, Virginia, had imaged mouse embryos every five minutes as they developed, yielding terabytes of data and a high-resolution developmental atlas[3] that has generated significant interest. "People are forever asking me, did you look at this tissue, did you look at that tissue?" So she looked for a way to share all those terabytes.

The journal offered only gigabytes of space, much less than McDole needed. ("Oh, gigabytes," she scoffs, "gigabytes are cute.") So she uploaded the atlas to the IDR, a free service developed by Swedlow and his colleagues. The data transfer took the better part of a week, she says. But now, anyone with a web browser can scroll through her data set, find their favourite tissues, or compare their results with hers. McDole herself often uses the IDR at conferences, to show colleagues data she doesn't carry on her laptop.

Such databases offer more than a storage location, says Jan Ellenberg, a cell and molecular biologist at EMBL in Heidelberg, Germany, and researchers shouldn't simply drop their data sets into small, project-specific archives or generic cloud storage. "Just dumping the data somewhere doesn't mean people can use it," Ellenberg explains. "You need to organize the data, you need to annotate it, and curate it." Browsers of McDole's data set, for instance, can scan the metadata to find out information such as the strain of mice she used and the specific fluorescent labels she imaged.

Patrick Combes, global technical leader for health care and life sciences at Amazon Web Services in Seattle, agrees. "Storing a data set on Amazon doesn't automatically enhance it," he says. But if scientists handle processing, curation and annotation, Amazon can be a secure, reliable data host, he says. It already houses several widely used resources, including raw data from the Allen Brain Observatory and NeuroData.

Researchers can typically upload their data to life-science image databases at no cost, because storage, curation and maintenance are often funded by grants or other benefactors. Shuichi Onami, a developmental biologist at the RIKEN Center for Biosystems Dynamics Research in Kobe who founded the SSBD database, obtained funding from institutions including RIKEN; the Japan Science and Technology Agency; and the nation's Ministry of Education, Culture, Sports, Science and Technology. The database is "completely free" to the user, says Onami. Now he is expanding it beyond developmental biology, to include any biological data set that contains spatiotemporal information, as well as static images taken with state-of-the-art technologies.

> **"We do all of the boring infrastructure to make sure those data persist."**

It's also generally free to download data sets, and often to reuse and republish them: repositories frequently use Creative Commons licences that make availability transparent.

Pick and choose

Databases differ in the sizes of files they will accept, whether images must be linked to a published study, and their research focus. If your scientific community already has a specialized data house, Vogelstein recommends using that.

But there are general repositories. Figshare, for instance, accepts any kind of data, up to 5 gigabytes per file, for free. It can sometimes raise the limit, says founder Mark Hahnel — the biggest Figshare data set measures in terabytes. (Figshare is owned by Digital Science, a firm operated by the Holtzbrinck Publishing Group, which has a share in *Nature*'s publisher, Springer Nature.) Other free, catch-all type services include Zenodo and Dryad.

Figshare also has contracts with universities, funders and publishers (including Springer Nature), which pay an annual fee for extra benefits. Last year, it set up a repository of data from research funded by the US National Institutes of Health (NIH), which now expects its grant recipients to make their results freely available. The site is meant for data that don't fit neatly into subject-specific banks, and currently hosts dozens of data sets. Unlike with the standard Figshare service, the NIH has control over the repository: what kinds of content are allowed, for example, and what kinds of metadata are required. NIH grantees benefit from assistance with metadata for their submissions, among other features.

The IDR databases — there is one for images of cells and one for tissues — are tightly curated, says Swedlow. He and the other curators seek reference data sets linked to publications, such as results from large screening studies that would be of use to a wide audience. They ensure that the data are properly formatted and annotated with relevant metadata, such as information on the microscope used and experimental treatments applied.

Last July, EMBL-EBI announced a service called the BioImage Archive, which will host both the IDR and EMPIAR, as well as the more general BioStudies database. The institute will support further curated, community-specific databases in future, says Jo McEntyre, associate director for services at EMBL-EBI. With support from EMBL and the funding agency UK Research and Innovation, the BioImage Archive will be maintained for "as long as it's scientifically useful", she promises. Figshare, Hahnel says, "will persist forever" — although he admits the contract guarantees only a decade. "We do all of the boring infrastructure to make sure those data persist," he says.

Other people's data

These services make it easier to find, share and store big data sets. But as with DNA and protein databases, the hope is that image-surfers will find new science in others' data.

Demonstrating this potential, Swedlow and his colleagues combed images from three separate studies of cell elongation in the IDR. Two were from the human cancer-cell line HeLa; one was in fission yeast; all three imaged cells missing a variety of genes. "Each study gets different results, but they're related," says Swedlow. Together, these studies allowed him and his team to identify a larger, more complete network of genes involved in elongation than they could get from any one data set alone[4].

A study[5] posted on the arXiv preprint server last year reports that of almost 532,000 journal articles published by PLOS and BioMed-Central, those that linked to a data repository had up to a 25% higher citation impact than those that didn't.

With time to mature, image databases could yield more than just one-off discoveries, Swedlow says. After all, bioinformatics itself grew out of DNA archives.

"Hopefully," he says, "we end up stimulating the development of whole fields."

Amber Dance is a freelance science journalist near Los Angeles, California.

1. Zivanov, J., Nakane, T. & Scheres, S. H. W. *IUCrJ* https://doi.org/10.1107/S2052252520000081 (2020).
2. Dorkenwald, S. *et al.* Preprint at bioRxiv https://doi.org/10.1101/2019.12.29.890319 (2019).
3. McDole, K. *et al. Cell* **175**, 859–876 (2018).
4. Williams, E. *et al. Nature Meth.* **14**, 775–781 (2017).
5. Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K. & McGillivray, B. Preprint at https://arxiv.org/abs/1907.02565 (2019).