ILLUSTRATION BY THE PROJECT TWINS

HOW TO GET MICROSCOPES TO SPEAK THE SAME LANGUAGE

A plethora of standards mean shareable and verifiable microscopy data often get lost in translation. Biologists are working on a solution. **By Michael Brooks**

Many of us can relate to characters in films or fiction; Jason Swedlow sees himself in the stick figures of the online comic strip *xkcd*.

In a strip published in 2011, cartoonist Randall Munroe pokes fun at people's inability to develop a universal standard for, say, electrical outlets, printer-paper dimensions or TV remote-control signals. From 14 competing standards in the opening panel, the desire to create a 'universal' standard inevitably just adds one more to the stack (https://xkcd.com/927).

"That comic is shown at almost every presentation I attend," says Swedlow, a cell biologist at the University of Dundee, UK.

Swedlow has been working for two decades to standardize image formats for biological microscopy data. During that time, the number of standard file formats in the field has proliferated to around 160. Now, thanks to a project in which Swedlow has a leading role, there is one more. "For 20 years we've been trying to solve this file-format problem," he says. "And how are we going to solve it? Come up with a new one."

Swedlow can't help but laugh. But he and his colleagues are aiming to prove Munroe's cartoon wrong. Josh Moore, one of Swedlow's collaborators, reckons that they can shrink that file-format mountain down to a small handful. "I feel like that's something that's manageable from our side," says Moore, senior research data-management officer at German BioImaging, a network for the nation's microscopists and bioimage analysts based in Konstanz, Germany.

"Our side" is OME-Zarr, a blend of two projects. The first is the Open Microscopy Environment (OME), which Swedlow started in 2002 to develop an open-source specification for biological microscopy data. Zarr is a newer creation: a method for optimizing how large data arrays are stored in, and downloaded from, the cloud. In 2021, Moore and his colleagues reported the first specification for OME and Zarr to work together as a next-generation file format (NGFF) for bioimaging (J. Moore *et al. Nature Methods* **18**, 1496–1498; 2021). This year, OME-Zarr launched as a fully-fledged option for biologists to store data, with support

from dozens of specially developed tools and programming libraries. Now the real test begins: can team OME-Zarr persuade everyone involved in bioimaging that speaking the same data language is the path to microscopy utopia?

Mountains of data

Modern microscopes create mountains of data, with researchers pushing the instruments to produce images at ever higher spatial resolution, in ever more colours and for longer periods. Each pixel must be labelled with metadata, such as illumination level, its 3D position, the scale, the sample type and how the sample was prepared. Between raw data and metadata, a lab can easily produce a hard-drive's worth of information in a day.

That in itself is not a huge problem: data storage is getting cheaper all the time. But — and this is where the *xkcd* analogy comes in — every microscope manufacturer formats its metadata differently. This is also true of the many do-it-yourself systems made in individual labs. What's more, reading the metadata tags for each manufacturer's image files often requires software created specifically for that system. In an era when researchers are striving to make their data findable, accessible, interoperable and reusable (also known as FAIR), this is a huge problem.

Take Katrín Möller's experience. As part of her graduate research at the University of Zurich in Switzerland, Möller imaged cells called microglia in living zebrafish brains. "A single session could produce a terabyte of data," she says.

The metadata were an essential part of that information. "A lot of the things that I did in that project involved measuring distances of travel or where things were located in 3D space," says Möller, who earned her PhD in 2022 and is now a postdoc at the University of Iceland in Reykjavik. "I had to capture both the time metadata and the spatial metadata: locations and pixel size, and which pixel it is. All this metadata had to be stored in the raw data, otherwise I'd have to write it down for every single data set."

Möller test-drove three microscopes, and all were capable of formatting and outputting the data for storage. But each one did it in a different way, and none of them was compatible with the software that she used to process and analyse the data. In the end, Möller resorted to converting the output of her chosen microscope into TIFF files by hand. "Sometimes I would spend the whole day converting things to a usable format," she says.

Moore recalls one biologist who was studying chicken embryos and needed to measure a particular angle at every frame of a 72-hour experiment. She did it by hand, logging the metadata in Microsoft Excel. "She was willing to suffer because she wanted to do her science," Moore says. "The formats problem is just this thing that we tolerate."

And yet people don't have to. Möller sped up her conversions by writing macros to handle most of her processing work, and larger institutions can write their own software. But those are siloed solutions — customized to the researchers for whom they were written and unavailable to the wider field. They're not even guaranteed to work if the manufacturer issues a new release of its software. "Versioning is a big problem," Moore says.

"Sometimes I would spend the whole day converting things to a usable format."

Few manufacturers support old versions of their software, Moore explains — they say they lack the resources. But the team behind OME has to think bigger: it aims to support everything that biologists might be using or have ever used, because the information in old files still needs to be accessible.

It also has to be trustworthy. Fraud investigators such as the Office of Research Integrity (ORI) in Rockville, Maryland, have welcomed efforts to open microscope vendors' proprietary file formats to everyone, for instance, because it simplifies its work. Although initial investigations of alleged research fraud are typically carried out by just looking at the images, having access to the files themselves is essential, says Chad McCormick, a scientist investigator at the ORI. "For microscopy images, it is important to show that there are unique source files and that these files, or any subsequent 2D representation of these files, do not contain manipulations," he says.

Greta Sharpe, research-integrity specialist at Springer Nature, which publishes *Nature*, says that this can be far from straightforward. "Authors sometimes provide low-quality images with no relevant metadata as their raw data," she explains. (*Nature*'s journalism team is editorially independent of its publisher.)

That matters because if two images look similar, it's useful to look deeper. If the files were created within a few seconds of each other, for example, it's more likely that they originated from the same sample, Sharpe says. Missing metadata might be the result of an innocent attempt to save time and effort, but it could also be a red flag for images generated by artificial intelligence.

Culture shift

Layered on all this is another complication: the ephemeral and remote-data-storage solution known as the cloud.

Your standard personal computer stores files that contain a 'file pointer', a digital cursor that points to the data you're interested in. By moving that cursor, researchers can pull data from anywhere in the file — allowing random access.

The cloud, however, treats data as a single unstructured entity that is either downloaded in its entirety or not — called 'object storage'. That's fine if your file is a PDF document or a holiday photo. If it's a terabyte-sized data set, it's like dropping a suitcase on kitchen scales. "Object stores are dumb!" Swedlow says. But, with researchers flocking to put their data in the cloud, he and his colleagues had no choice but to adjust.

Zarr provides a generic method for storing and accessing data arrays, such as the succession of binary digits that make up a stack of image files. It breaks the arrays into chunks that can be compressed in a way that retains all the information but still allows fast reading of, and writing to, the file.

For microscopy data, Zarr stores neighbouring pixels in the same chunk, so that they arrive together when downloaded. They also arrive quickly, because each chunk can be compressed without losing any information. The user can set the size of the chunk, too, allowing optimization of file size, number of files, level of resolution, and read and write speeds.

David Feng, who leads scientific computing at the Allen Institute for Neural Dynamics in Seattle, Washington, is part of a research team that is using OME-Zarr to help power a microscopy system called expansion-assisted selective plane illumination microscopy (ExA-SPIM). With the ability to image an entire mouse brain at nanoscale resolution, the system can produce around 100 terabytes a day. The only way to handle that much data is to get it into the cloud as fast as possible, Feng says. After a lot of benchmarking, the team chose to do that using OME-Zarr. "For users, it's very easy to only download the data you want to download," he explains: "You just grab the little chunk that's of interest."

First steps

The OME standard adds to this convenience by providing a multiscale representation for microscope data, similar to how Google Maps lets you see the world at any length scale without overwhelming your mobile phone's processor. "Rather than having a gigantic 100 terabyte file, you have different levels of lower resolution: tiers of a pyramid that you can access depending on what you want to see," Feng says.

That flexibility is particularly valuable for biologists, because it allows collaboration between separate groups by making it possible for them to view the file, says Beth Cimini, who is the associate director for bioimage analysis at the Broad Institute of MIT and Harvard in Cambridge, Massachusetts. "Step one of someone being able to use your data is them being able to actually open your data," she says.

That said, there is a step zero: persuading

biologists to think of data sharing as more than just a recipe for image theft. "People are always asking us how they can keep track of how their data is being used if they share it," says Shuichi Onami, who leads the developmental dynamics laboratory at the Riken Center for Biosystems Dynamics Research in Kobe, Japan.

Still, despite researcher reservations, Onami is convinced that a cultural shift away from data protectionism is happening, in part, thanks to pressure from publishers. And that external pressure will continue to be essential, adds Catherine Maclachlan, a senior laboratory research scientist at the Francis Crick Institute in London, because scientists trust the formats that they know. "When you've spent ages perfecting and collecting your data, you don't want to risk anything. Change tends to come only when you really have to change – such as when a journal says it has to be in this particular format."

Thanks to projects such as OME, conversion software is readily available.

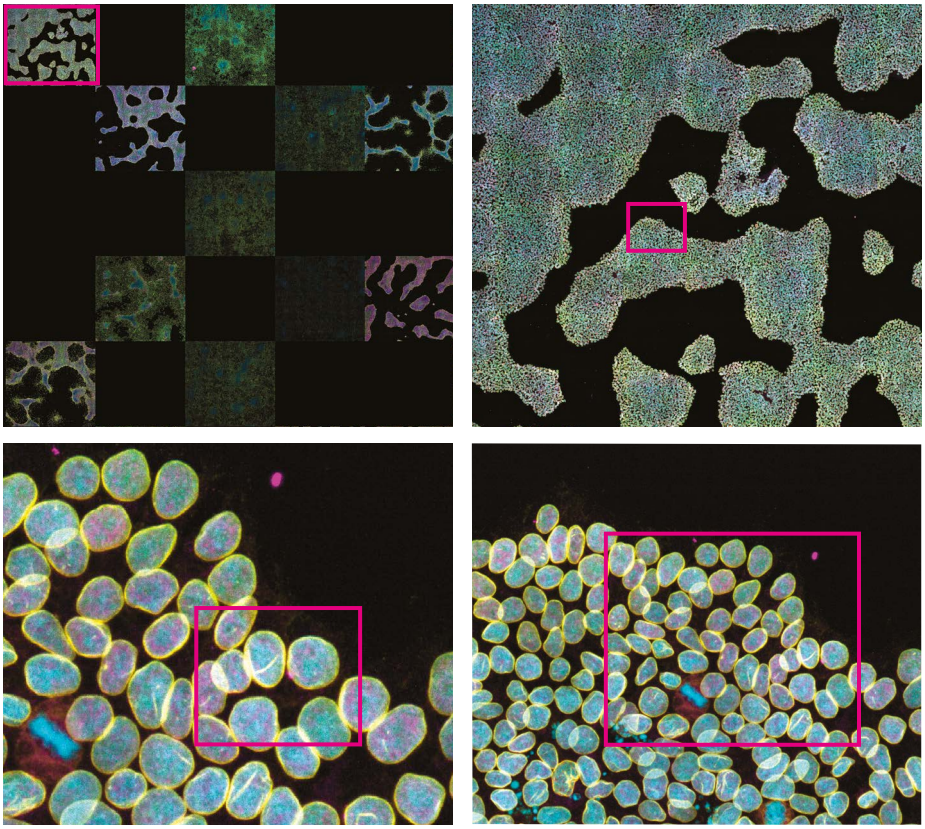
But growing adoption of OME-Zarr remains an uphill struggle, says Maclachlan's colleague Martin Jones, who is deputy head of microscopy prototyping in the electron microscopy science and technology platform at the Francis Crick Institute. After all, biologists have enough to do without having to learn to handle new forms of data. And the Zarr format can be a little daunting, he admits. Biologists are used to being able to convert a standard image file into one that can be opened in a spreadsheet program, in which they can see data such as pixel sizes and intensities represented as numbers. Open a Zarr archive and you'll just see a seemingly endless set of nested folders. "There's no way you can know what that is," he says.

The other issue is that file formats are a bit dull. "I gave a talk once," Moore says, "and a principal investigator asked, 'Do I actually need to know any of that? Do I need to engage with this?'"

At the moment, Moore says, the answer is yes, because NGFF enthusiasts need biologists on board with the effort to get microscope vendors to output a common, agreed format from their instruments.

Vendor perspectives

It would be easy to lay blame at the feet of the various microscope manufacturers. But Matthias Genenger, a product manager at microscope vendor Evident (formerly Olympus) in Münster, Germany, says that the diversity of file formats is inevitable because of commercial competition. Although his company has been building compatibility with NGFFs such as OME-Zarr for some time, open-source software doesn't always cover all of a microscope's functionality. As manufacturers improve their microscopes, open file formats will inevitably lag behind. "Some of our products are very specific, and the open or generic file format does not give us all the flexibility we need to integrate the maximum performance into these products," Genenger says.



J. MOORE *ET AL. HISTOCHEM. CELL BIOL.* **160**, 223–251 (2023)

The OME-Zarr file format lets users select data in a multi-resolution image of cells (pink square, upper left) and zoom in (clockwise from upper right) while accessing only the pixels they need.

Furthermore, there's little incentive for manufacturers to change, advocates concede. "We have to make it worth their while," says Cimini. "If we want them to abandon these formats that they spent time and effort making, we have to show them that there's some value in it for them."

> **"A principal investigator asked, 'Do I actually need to know any of that?'"**

Biologists have to put their house in order, too. OME-Zarr isn't the only open-source game in town. One alternative is N5, a Zarr-like format that tends to be favoured by people who process data using Java-based software tools, such as Fiji (OME-Zarr is easier to use with the Python programming language). And the HDF5 format is better for those who share data by copying or downloading files, says John Bogovic, a machine-learning researcher at the Howard Hughes Medical Institute's Janelia research campus in Ashburn, Virginia. Manufacturer formats are useful, too. "Although proprietary, Zeiss's CZI is decently open, useful and has a big user base, because Zeiss hardware uses it," Bogovic says, referring to the German microscope manufacturer.

There is no consensus yet on exactly which bioimage file format – or set of them – vendors should adopt, but the situation needs resolving, Moore says. "It is incumbent on the wider community to say 'here's what we want you to do' and then everyone can play along."

Antje Keppler, director of the Euro-BioImaging Bio-Hub at the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany, agrees. "The manufacturers are quite active and eager," she says. "In my view, they would be on board as soon as the community can lead the way."

This brings us back to the issue of data formatting, which – for some people – can be a bit of a bore. Swedlow says he can understand why not every biologist shares his passion for getting to grips with bioimaging file formats. "It's not a very interesting problem," he admits. Moore agrees. "Let's be honest, when this whole topic disappears, that's going to be a good thing."

But not, perhaps, for Munroe's page views. After a long conversation at the Francis Crick Institute, Jones has one final thing to share with *Nature* about the topic of bioimaging file formats. "Are you familiar with the *xkcd* comics?" he asks.

Michael Brooks is a science writer in Lewes, UK.