

Digital Preservation at Oxford and Cambridge

A collaborative research project to evaluate and provide sustainable recommendations for our digital preservation programmes

Preserving research – update from the Cambridge Technical Fellow

Posted on [13 June, 2017](#) by [Sarah](#)

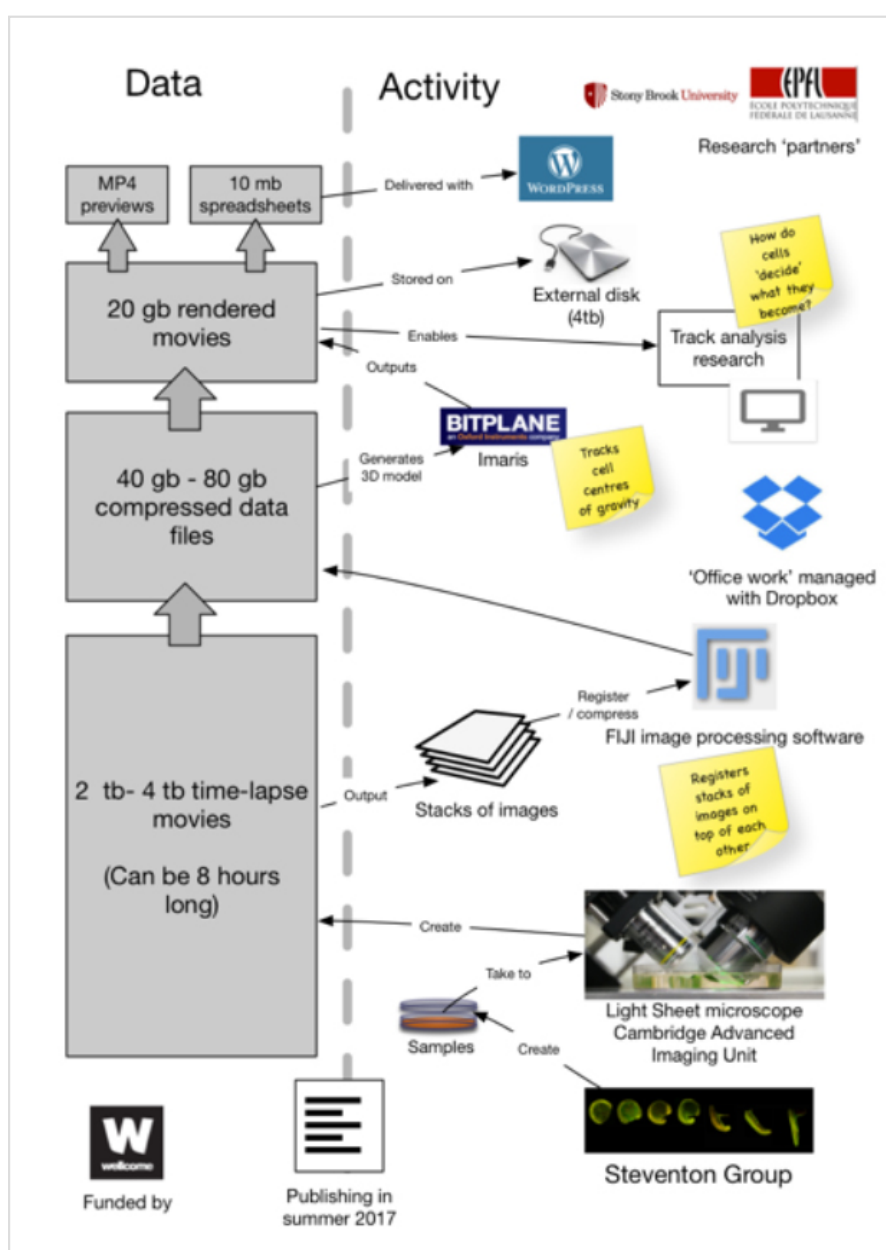
Cambridge's Technical Fellow, Dave, discusses some of the challenges and questions around preserving 'research output' at Cambridge University Library.

One of the types of content we've been analysing as part of our initial content survey has been labelled 'research output'. We knew this was a catch-all term, but (according to the categories in Cambridge's [Apollo Repository](#)), 'research output' potentially covers: "Articles, Audio Files, Books or Book Chapters, Chemical Structures, Conference Objects, Datasets, Images, Learning Objects, Manuscripts, Maps, Preprints, Presentations, Reports, Software, Theses, Videos, Web Pages, and Working Papers". Oh – and of course, "Other". Quite a bundle of complexity to hide behind one simple 'research output' label.

One of the categories in particular, 'Dataset', zooms the fractal of complexity in one step further. So far, we've only spoken in-depth to a small set of scientists (though our participation on Cambridge's [Research Data Management Project Group](#) means we have a great network of people to call on). However, both meetings we've had

indicate that 'Datasets' are a whole new Pandora's box of complicated management, storage and preservation challenges.

However – if we pull back from the complexity a little, things start to clarify. One of the scientists we spoke to (Ben Steventon at the [Steventon Group](#)) presented a very clear picture of how his research 'tiered' the data his team produced, from 2-4 terabyte outputs from a Light Sheet Microscope (at the [Cambridge Advanced Imaging Centre](#)) via two intermediate layers of compression and modelling, to 'delivery' files only megabytes in size. One aspect of the challenge of preserving such research then, would seem to be one of tiering preservation storage media to match the research design.



(I believe our colleagues at the JISC, who Cambridge are working with on the [Research Data Management Shared Service Pilot Project](#), may be way ahead of us on this...)

Of course, tiering storage is only one part of the preservation problem for research data: the same issues of acquisition and retention that have always been part of archiving still apply... But that's perhaps where the 'delivery' layer of the Steventon Group's research design starts to play a role. In 50 or 100 years' time, which sets of the research data might people still be interested in? It's obviously very hard to tell, but perhaps it's more likely to be the research that underpins the *key* model: the major finding?

Reaction to the 'delivered research' (which included papers, presentations and perhaps three or four more from the list above) plays a big role, here. Will we keep all 4TBs from every Light Sheet session ever conducted, for the entirety of a five or ten-year project? Unlikely, I'd say. But could we store (somewhere cold, slow and cheap) the 4TBs from *the* experiment that confirmed the major finding?

That sounds a bit more within the realms of possibility, mostly because it feels as if there might be a chance that someone might want to work with it again in 50 years' time. One aspect of modern-day research that makes me feel this might be true is the complexity of the dependencies between pieces of modern science, and the software it uses in particular. ([Blender](#), for example, or [Fiji](#)). One could be pessimistic here and paint a negative scenario of 'what if a major bug is found in one of those apps, that calls into question the science 'above it in the chain'. But there's an optimistic view, here, too... What if someone comes up with an entirely new, more effective analysis method that replaces something current science depends on? Might there not be value in pulling the data from old experiments 'out of the archive' and re-running them with the new kit? What would we find?

We'll be able to address some of these questions in a bit more detail later in the project. However, one of the more obvious things talking to scientists has revealed is that many of them seem to have large collections of images that need careful management. That seems quite relevant to some of the more 'close to home' issues we're looking at right now in The Library.

SHARE THIS:



This entry was posted in [digital curation](#), [digital preservation](#), [research data management](#) by [Sarah](#). Bookmark the [permalink](#)

[\[http://www.dpoc.ac.uk/2017/06/13/preserving-research-update-from-the-cambridge-technical-fellow/\]](http://www.dpoc.ac.uk/2017/06/13/preserving-research-update-from-the-cambridge-technical-fellow/) .

About Sarah

Digital Preservation Specialist - Outreach and Training:
Bodleian Libraries, Oxford University

[View all posts by Sarah](#) →

ONE THOUGHT ON “PRESERVING RESEARCH – UPDATE FROM THE CAMBRIDGE
TECHNICAL FELLOW”

ehalvarsson

on **14 June, 2017 at 11:13** said:

Thanks for this post Dave!

I also find it hard to get my head around the large scale outputs from these tools. However, I do wonder if only retaining data from key findings is really the right approach. Experiments which do not underpin ‘main’ findings may still be of value in that they confirm what is not there.

Depending on the area of study, mapping these are also a type of finding which may avoid duplication of effort for other researchers.

This site uses Akismet to reduce spam. [Learn how your comment data is processed.](#)