

Digital Preservation at Oxford and Cambridge

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

Putting ‘stuff’ in ‘context’: deep thoughts triggered by PASIG 2017

Posted on [12 October, 2017](#) by [Dave Gerrard](#)

Cambridge Technical Fellow, Dave, delves a bit deeper into what PASIG 2017 talks really got him thinking further about digital preservation and the complexity of it.

After a year of studying digital preservation, my thoughts are starting to coalesce, and the [presentations at PASIG 2017](#) certainly helped that. (I’ve already discussed [what I thought were the most important talks](#), so the ones below some that stimulated me about preservation in particular)...

The one that matched my current thoughts on digital preservation generally was John Sheridan’s [Creating and sustaining a disruptive digital archive](#). It was similar to [another previous blog post](#), and to chats with fellow Fellow Lee too (some of which he’s captured in a [blog post for the Digital Preservation Coalition](#))... I.e.: computing’s ‘paper paradigm’ makes little sense in relation to preservation, hierarchical / neat information structures don’t hold together as well digitally, we’re going to need to compute across the whole archive, and, well, ‘digital objects’ just aren’t really material ‘objects’, are they?

An issue with thinking about digital 'stuff' too much in terms of tangible objects is that opportunities related to the fact the 'stuff' is digital can be missed. Matt Zumwalt highlighted one such opportunity in [Data together: Communities & institutions using decentralized technologies to make a better web](#) when he introduced 'content addressing': using cryptographic hashing and Directed Acyclic Graphs (in this case, information networks that record content changing as time progresses) to manage many copies of 'stuff' robustly.

This addresses some of the complexities of preserving digital 'stuff', but perhaps thinking in terms of 'copies', and not 'branches' or 'forks' is an over simplification? Precisely because digital 'stuff' is rarely static, all 'copies' have the potential to deviate from the 'parent' or 'master' copy. What's the 'version of true record' in all this? Perhaps there isn't one? Matt referred to 'immutable data structures', but the concept of 'immutability' only really holds if we think it's possible for data to ever be completely separated from its informational context, because the *information* does change, constantly. (Hold that thought).

Switching topics, fellow Polonsky Somaya often tries to warn me just how complicated working with technical metadata can get. Well, the pennies dropped further during [Managing digital preservation metadata at Sound and Vision: A case on matching OAIS and PREMIS with the DPX file format](#) from Annemieke De Jong and Josefiën Schuurman. Space precludes going into the same level of detail they did regarding building a Preservation Metadata Dictionary (PMD) about *just one, 'relatively' simple file format* – but let's say, well, it's really complicated. (They've [blogged about it](#) and [the whole PMD](#) is online too). The conclusion: preserving files properly means drilling down deep into their formats, but it also got me thinking – shouldn't the essence of a 'preservation file format' be its simplicity?

The need for greater simplicity in preservation was further emphasised by Mathieu Giannecchini's [The Eclair Archive cinema heritage use case: Rising to the challenges of complex formats at large scale](#). Again – space precludes me from getting into detail, but the key takeaway was that Mathieu has 2 million reels of film to preserve using the Digital Cinema Distribution Master (DCDM) format, and after lots of good work, he's optimised the process to preserve 8tb a day, (with a target of 15tb). Now, we don't know how much film is on each reel, but assuming a (likely over-) estimate of 10 minutes per reel, that's roughly 180,000 films of 1 hour 50 mins

in length. Based on [Mathieu's own figures](#), it's going to take many decades, perhaps even a few hundred years, to get through all 2 million reels... So further, major optimisations are required, and I suspect DCDM (a format with a 155-page spec, which relies on TIFF, a format with a 122-page spec) might be one of the bottlenecks.

Of course, the trade-off with simplifying formats is that data will likely be 'decontextualised', so there must be a robust method for linking data back to context... Thoughts on this were triggered by [Developing and applying principles for discovery and access for the UK Data Service](#) by Katherine McNeill from the UK Data Archive, as Katherine discussed production of a next-generation access system based on a linked-data model with which, theoretically, single cells' worth of data could be retrieved from research datasets.

Again – space precludes entering into the whole debate around the process of re-using data stripped of original context... Mauthner and Parry [illustrate the two contrary sides](#) well, and furthermore argue that merely entertaining the possibility of decontextualising data indicates a certain 'foundational' way of thinking that might be invalid from the start? This is where I link to [William Kilbride's excellent DPC blog post from a few months ago](#)...

William's PASIG talk [Sustainable digital futures](#) was also one of two that got closer to what we know are the root of the preservation problem; economics. The other was [Aging of digital: Managed services for digital continuity](#) by Natasa Milic-Frayling, which flagged-up the current "imbalance in control and empowerment" between tech providers and content producers / owners / curators, an imbalance that means tech firms can effectively doom our digital 'stuff' into obsolescence, and we have to suck it up.

I think this imbalance in part exists because there's too much *technical context* related to data, because it's generally in the tech providers' interests to bloat data formats to match the USPs of their software. So, is a pure 'preservation format' one in which the technical context of the data is generalised to the point where all that's left is commonly-understood mathematics? Is that even possible? Do we *really* need 122-page specs to explain how raster image data is stored? (It's just an N-dimensional array of pixel values..., isn't it...?) I think perhaps we *don't* need all the complexity – at the data storage level at least. Though I'm only guessing at this stage: much more research required.

SHARE THIS:



Share

This entry was posted in [digital preservation](#), [PASIG2017](#), [research data management](#), [technology](#) by [Dave Gerrard](#).

Bookmark the [permalink](#)

[\[http://www.dpoc.ac.uk/2017/10/12/putting-stuff-in-context-deep-thoughts-triggered-by-pasig-2017/\]](http://www.dpoc.ac.uk/2017/10/12/putting-stuff-in-context-deep-thoughts-triggered-by-pasig-2017/) .

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