

Digital Preservation at Oxford and Cambridge

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

On the core concepts of digital preservation

Posted on [4 November, 2016](#) by [Dave Gerrard](#)

Cambridge's Technical Fellow, Dave Gerrard, shares his learning on digital preservation from the PASIG 2016. As a newcomer to digital preservation, he is sharing his insights as he learns them.

As a relative newbie to Digital Preservation, attending [PASIG 2016](#) was an important step towards getting a picture of the state of the art in digital preservation. One of the most important things for a technician to do when entering a new domain is to get a high-level view of the overall landscape, and build up an understanding of some of the overarching concepts, and last week's PASIG conference provided a great opportunity to do this.

So this post is about some of those central overarching data preservation concepts, and how they might, or might not, map onto 'real-world' archives and archiving. I should also warn you that I'm going to be posing as many questions as answers here: it's early days for our Polonsky project, after all, so we're all still definitely in the 'asking' phase. (Feel free to answer, of course!) I'll also be contrasting two particular presentations that were delivered at PASIG, which at first glance have little in common, but which I thought actually made the same point from completely different perspectives.

Perhaps the most obvious, key concept in digital preservation is 'the archive': a place where one deposits (or donates) things of value to be stored and preserved for the long term. This concept inevitably influences a lot of the theory and activity related to preserving digital resources, but is there really a direct mapping between how one would preserve 'real' objects, in a 'bricks and mortar' archive, and the digital domain? The answer appears to be 'yes and no': in certain areas (perhaps related to concepts such as acquiring resources and storing them, for example) it seems productive to think in broadly 'real-world' terms. Other 'real-world' concepts may be problematic when applied directly to digital preservation, however.

For example, my fellow Fellows will tell you that I take particular issue with the word 'managing': a term which in digital preservation seems to be used (at least by some people) to describe a particular small set of technical activities related to checking that digital files are still usable in the long-term. ('Managing' was used in this context in at least one PASIG presentation). One of the keys to working effectively with Information Systems is to get one's terminology right, and in particular, to group together and talk about parts of a system *that are on the same conceptual level*. I.e. don't muddle your levels of detail, particularly when modelling things. 'Managing' to me is a generic, high-level concept, which could mean anything from 'making sure files are still usable' to 'ensuring public-facing staff answer the phone within five rings' or even 'making sure the staff kitchen is kept clean'. So I'm afraid that I think it's an entirely inappropriate word to describe a very specific set of technical activities.

The trouble is, most of the other words we've considered for describing the process of 'keeping files usable' are similarly 'higher-level' concepts... One obvious one (preservation) once again applies to much more of the overall process, and so do many of its synonyms ('stewardship', 'keeping custody of', etc...) So these are all good terms *at that high level of abstraction*, but they're for describing the big picture, not the details. Another term that is more specific, 'fixity checking', is maybe a bit too much like jargon...

(We're still working on this: answers below please!) But the key point is: until one understands a concept well enough to be able to describe it in relatively simple terms, that make sense and fit together logically, building an information system and marshalling the related technology is always going to be tough.

Perhaps the PASIG topic that highlighted the biggest difference between ‘real world’ archiving and digital preservation the most, however, was discussion regarding the increased rate at which preserved digital resources can be ‘touched’ by outside forces. Obviously, nobody stores things in a ‘real-world’ archive in the expectation that they will never be looked at again (do they?), but in the digital realm, there are potentially many more opportunities for resources to be linked directly to the knowledge and information that builds upon them.

This is where the two contrasting presentations came in. The first was [Scholarly workflow integration: The key to increasing reproducibility and preservation efficacy](#), by Jeffrey Spies (@JeffSpies) from the [Center for Open Science](#). Jeffrey clarified exactly how digital preservation in a research data management context can highlight, explicitly, how a given piece of research builds upon what went before, by enabling direct linking to the publications, and (increasingly) to the raw data of peers working in the same field. Once digital research outputs and data are preserved, they are available to be linked to, reliably, in a manner that brings into play entirely new opportunities for archived research that never existed in the ‘real world’ of paper archives. Thus enabling the ‘discovery’ of preserved digital resources is not just about ensuring that resources are well-indexed and searchable, it’s about adding new layers of meaning and interpretation as future scholars use them in their own work. This in turn indicates how digital preservation is a function that is entirely integral to the (cyclical) research process – a situation which is well-illustrated in the 20th slide from Jeffrey’s presentation (if you download it – Figshare doesn’t seem to handle the animation in the slide too well – which sounds like a preservation issue in itself...).

By contrast, [Symmetrical Archiving with Webrecorder](#), a talk by Dragan Espenschied (@despens), was at first glance completely unrelated to the topic of how preserved digital resources might have a greater chance of changing as time passes than their ‘real-world’ counterparts. Dragan was demonstrating the [Webrecorder](#) tool for capturing online works of art by recording visits to those works through a browser, and it was during the discussion afterwards that the question was asked: “how do you know that everything has been recorded ‘properly’ and nothing has been missed?”

For me, this question (and Dragan’s answer) struck at the very heart of the same issue. The answer was that each recording is a different object in itself, as the interpretation of the person recording

the artwork is an integral part of the object. In fact, Dragan's exact answer contained the phrase: "when an archivist adds an object to an archive, they create a new object"; the actual act of archiving changes an object's meaning and significance (potentially subtly, though not always) to an extent that it is *not the same object* once it has been preserved. Furthermore, the object's history and significance change once more with every visit to see it, and every time it is used as inspiration for a future piece of work.

Again – I'm a newbie, but I'm told by my fellow Fellows this situation is well understood in archiving and hence may be more of a revelation to me than most readers of this post. But what *has* changed is the way the digital realm gives us the opportunity not just to record how objects change as they're used and referred to, but also a chance to make the connections to new knowledge gained from use of digital objects *completely explicit* and part of the object itself.

This highlights the final point I want to make about two of the overarching concepts of 'real-world' archiving and preservation which PASIG indicated might not map cleanly onto digital preservation. The first is the concept of 'depositing'. According to Jeffrey Spies's model, the 'real world' research workflow of 'plan the research, collect and analyse the data, publish findings, gain recognition / significance in the research domain, and then finally deposit evidence of this ground-breaking research in an archive', simply no longer applies. In the new model, the initial 'deposit' is made at the point a key piece of data is first captured, or a key piece of analysis is created. Works in progress, early drafts, important communications, grey literature, as well as the final published output, are all candidates for preservation *at the point they are first created by the researchers*. digital preservation happens seamlessly in the background. The states of the 'preserved' objects change throughout.

The second is the concept of 'managing' (urgh!), or otherwise 'maintaining the status quo' of an object into the long-term future. In the digital realm, there doesn't need to be a 'status quo' – in fact there just isn't one. We can record when people search for objects, when they find them, when they cite them. We can record when preserved data is validated by attempts to reproduce experiments or re-used entirely in different contexts. We can note when people have been inspired to create new artworks based upon our previous efforts, or have interpreted the work we have preserved from entirely new perspectives. This is genuine preservation:

preservation that will help fit the knowledge we preserve today into the future picture. This opportunity would be much harder to realise when storing things in a 'real-world' archive, and we need to be careful to avoid thinking too much 'in real terms' if we are to make the most of it.

What do you think? Is it fruitful to try and map digital preservation onto real world concepts? Or does doing so put us at risk of missing core opportunities? Would moving too far away from 'real-world' archiving put us at risk of losing many important skills and ideas? Or does thinking about 'the digital data archive' in terms that are too like 'the real world' limit us from making important connections to our data in future?

Where does the best balance between 'real-world' concepts and digital preservation lie?


Name (required)

Email (required)

Comment (required)

Submit »

SHARE THIS:

 Share

This entry was posted in [digital preservation](#), [Information Systems](#), [modelling](#), [PASIG2016](#), [technology](#) by [Dave Gerrard](#). Bookmark the [permalink \[http://www.dpoc.ac.uk/2016/11/04/on-the-core-concepts-of-digital-preservation/\]](http://www.dpoc.ac.uk/2016/11/04/on-the-core-concepts-of-digital-preservation/) .

Comments are closed.