

Digital Preservation at Oxford and Cambridge

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

Six Priority Digital Preservation Demands

Posted on [13 July, 2017](#) by [somaya](#)

Somaya Langley, Cambridge Policy and Planning Fellow, talks about her top 6 demands for a digital preservation system.



— Photo: Blazej Mikula, Cambridge University Library

As a former user of one digital preservation system (Ex Libris' Rosetta), I have spent a few years frustrated by the gap between

what activities need to be done as part of a digital stewardship end-to-end workflow – including packaging and ingesting ‘information objects’ (files and associated metadata) – and the maturity level of digital preservation systems.

Digital Preservation Systems Review

At Cambridge, we are looking at different digital preservation systems and what each one can offer. This has involved talking to both vendors and users of systems.

When I’m asked about what my top digital preservation system current or future requirements are, it’s excruciatingly hard to limit myself to a handful of things. However, having previously been involved in a digital preservation system implementation project, there are some high-level takeaways from past experiences that remain with me.

Shortlist

Here’s the current list of my six top ‘digital preservation demands’ (aka user requirements):

Integration (with various other systems)

A digital preservation ‘system’ is only one cog in a wheel within a much larger machine; one piece of a much larger puzzle. There is an entire ‘digital ecosystem’ that this ‘system’ should exist within, and end-to-end digital stewardship workflows are of primary importance. The *right amount* of metadata and/or files should flow should flow from one system to another. We must also know where the ‘source of truth’ is for each bit.

Standards-based

This seems like a no-brainer. We work in *Library Land*. Libraries rely on standards. We also work with computers and other technologies that also require standard ways (protocols etc.) of communicating.

For files and metadata to flow from one system to another – whether via import, ingest, export, migration or an exit strategy from a system – we already spend a bunch of time creating mappings and crosswalks from one standard (or implementation of a standard) to another. If we don’t use (or fully implement) existing standards, this means we risk mangling data, context or meaning;

potentially losing or not capturing parts of the data; or just wasting a whole lot of time.

Error Handling (automated, prioritised)

There's more work to be done in managing digital materials than there are people to do it. Content creation is increasing at exponential rates, meanwhile the number of staff (with the right skills) just aren't. We have to be smart about how we work. This requires prioritisation.

We need to have smarter systems that help us. This includes helping to prioritise where we focus our effort. Digital preservation systems are increasingly incorporating new third-party tools. We need to know which tool reports each error and whether these errors are show-stoppers or not. (For example: is the content no longer renderable versus a small piece of non-critical descriptive metadata that is missing?) We have to accept that, for some errors, we will never get around to addressing them.

Reporting

We need to be able to report to different audiences. The different types of reporting classes include (but are not limited to):

1. *High-level reporting* – annual reports, monthly reports, reports to managers, projections, costings etc.)
2. *Collection and preservation management reporting* – reporting on successes and failures, overall system stats, rolling checksum verification etc.
3. *Reporting for preservation planning purposes* – based on preservation plans, we need to be able to identify subsections of our collection (configured around content types, context, file format and/or whatever other parameters we choose to use) and report on potential candidates that require some kind of preservation action.

Provenance

We need to best support – via metadata – where a file has come from. This, for want of a better approach, is currently being handled by the digital preservation community through documenting changes as Provenance Notes. Digital materials acquired into our collections are not just the files, they're also the metadata. (Hence, why I refer to them as 'information objects'.) When an 'information

object' has been bundled, and is ready to be ingested into a system, I think of it as becoming an 'information package'.

There's a lot of metadata (administrative, preservation, structural, technical) that appears along the path from an object's creation until the point at which it becomes an 'information package'. We need to ensure we're capturing and retaining the important components of this metadata. Those components we deem essential must travel alongside their associated files into a preservation system. (Not all files will have any or even the right metadata embedded within the file itself.) Standardised ways of handling information held in Provenance Notes (whether these are from 'outside of the system' or created by the digital preservation system) and event information so it can be interrogated and reported on is crucial.

Managing Access Rights

Facilitating access is not black and white. Collections are not simply 'open' or 'closed'. We have a myriad of ways that digital material is created and collected; we need to ensure we can provide access to this content in a variety of ways that support both the content and our users. This can include access from within an institution's building, via a dedicated computer terminal, online access to anyone in the world, mediated remote access, access to only subsets of a collection, support for embargo periods, ensuring we respect cultural sensitivities or provide access to only the metadata (perhaps as large datasets) and more.

We must set a goal of working towards providing access to our users in the many different (and new-ish) ways they actually want to use our content.

It's imperative to keep in mind the whole purpose of preserving digital materials is to be able to access them (in many varied ways). Provision of content 'viewers' and facilitating other modes of access (e.g. to large datasets of metadata) are essential.

Final note: I never said addressing these concerns was going to be easy. We need to factor each in and make iterative improvements, one step at a time.

SHARE THIS:



This entry was posted in [digital preservation](#), [digital stewardship](#), [technology](#), [tools](#), [Uncategorized](#) by [somaya](#). Bookmark the [permalink](#) [<http://www.dpoc.ac.uk/2017/07/13/sixdigitalpreservationdemands/>]

13 THOUGHTS ON “SIX PRIORITY DIGITAL PRESERVATION DEMANDS”

somaya
on [24 July, 2017 at 11:45](#) said:

Hi Adi,

You’ve mentioned an important point – while consolidating efforts is important, having different tools to cross-check results is essential.

Back in 2005/2006 I started running DROID, JHOVE and the NLNZ Metadata Extractor across digital files, particularly those that would’ve been considered ‘less common’ at the time. Then again in 2011/2012, in using various digital preservation tools on digital collection materials, I noticed that (in one or two cases) different tools were classing a file as ‘audio’ and the other tool would report as ‘video’. Without multiple tools, I wouldn’t have even known to look into this.

As a user who ‘pulls the levers’ of a digital preservation system, I do need to be able to verify results and check against other tools to see if there are reported differences etc. Only then can I be confident that any digital preservation processes I’m undertaking or triggering are okay.

somaya
on [24 July, 2017 at 11:39](#) said:

Hi Kelly,

Thanks so much for your response.

In regards to your comment: “Moreover, while you may find your six demands satisfactorily addressed in your institution with a unique combination of tools, policy, and staffing (breathe in), those solutions need to be generalized (breathe out) so all can benefit.” – Absolutely!

These are just the 6 things that I’ve noticed while being a user in cultural institutions and in discussing collection materials with my colleagues here at Cambridge... however as I mentioned, I could continue this list of needed things till the cows come home, and then some.

Indeed each institution is unique, and of course every organisation is going to need their requirements addressed differently. In the case of Cambridge, as such as decentralised university, the potential complexity of needs and a ‘solution’ to support acquisition, management, preservation and provision of access isn’t a small or simple thing.

Adi Alter
on [23 July, 2017 at 12:30](#) said:

Hi Dave,

Thanks for your interesting suggestions. I agree that there are ways to solve this in some sort of an abstraction layer, and in a way this is how Rosetta’s extendable plugin module is designed, but I believe that this will still require specific adaptation for the various tools out there, so having at least some agreement on the output of such tools can significantly ease the required integration work.

Adi

Adi Alter
on [23 July, 2017 at 12:24](#) said:

Hi Somaya,

I fully agree with what you're saying. I think that the work done by OPF with VeraPDF and JHOVE is a great start in this direction, and it should be expanded, of course pending resources and funding.

Also, as I have hinted in my original mail, it will be good if we'll see consolidated efforts here, as currently there is some waste with different tools doing similar things, though one can argue that the need for different tools is important in preservation in order to ensure cross-checking.

We had some interesting discussions on this topic in the recent OPF annual general meeting, and I personally see it as one of the most important topics that should be prioritized for the near future.

Adi

Kelly

on **21 July, 2017 at 20:48** said:

Hi Somaya,

I jumped up and shouted 'YES' (honest) when I read your phrase in your response to Adi: "I think we need to find a way of the digital preservation community shouldering the burden, rather than it being carried by a handful of individuals."

There is no question in my mind that digital preservation is best approached collaboratively not only for the obvious reason that we work in an under-resourced field where sharing the load with colleagues helps us not feel (too) overwhelmed but also because it's also a really exciting and interesting set of challenges that can be approached from many directions. My digital preservation ideal world is a space where we can all (librarians, archivists, data managers, vendors, IT specialists, researchers) openly provide our perspective, share successes and failures,

make suggestions for both specific tools, and contribute to overarching preservation philosophies. Moreover, while you may find your six demands satisfactorily addressed in your institution with a unique combination of tools, policy, and staffing (breathe in), those solutions need to be generalized (breathe out) so all can benefit.

If I've learned one thing at Artefactual in the few months I've worked here, it's that digital preservation is not a one-size fits all enterprise but, each institution-specific solution helps to build a corpus of expertise across our domain.

Thank you Somaya!

Kelly Stewart
Systems Archivist
Artefactual Systems

somaya
on [20 July, 2017 at 19:30](#) said:

Hi Adi,

Thanks for your response – I appreciate you taking the time to feed back.

To me, there is a bigger underlying issue. I think the biggest challenge regarding the development of the various digital preservation tools created by the digital preservation community is the fact that much of this development is done by different individuals, for the most part, during their own personal time.

There is a lot of free-work volunteer effort that takes place to develop these tools that we all rely on (in different ways) To me, this isn't fully acknowledged by the digital preservation community as a whole. Without some of these tools, none of us would be able to do our jobs (not just the digital preservation systems). The people developing digital preservation tools do what they can, with the skills they have

and (often at night) produce what they need to help both themselves and their colleagues to do their day-jobs.

Given we work in GLAM/IT/information management/university sectors etc., it would be ideal if we're able to make incremental steps to standardising more. I do really believe this.

What I'm concerned about is that tools critical to a whole international community are the responsibility of one individual, one small group of people or one organisation. People work in this field because they are committed to the belief that others in our society should have access to culture and information. People developing digital preservation tools do feel personally responsible for supporting this endeavour. As a result, they put in considerable amounts of their personal time and energy into furthering this goal. I think we need to find a way of the digital preservation community shouldering the burden, rather than it being carried by a handful of individuals.

While I don't necessarily have the answers, I think there is a collaborative strategic approach needed here.

Dave Gerrard
on [20 July, 2017 at 15:22](#) said:

Hi Adi.

An alternative approach to trying to get people to agree on standard APIs is to lean on the array of integration design patterns commonly used in software to get different systems to co-operate with each other. The adaptor or plugin pattern is one commonly-used approach, for instance – it lets you translate from the third-party software's native data model to your local one. And there's a thing called a 'canonical data model' which basically formalises the local system's model so its easier to translate new pieces of third-party software as they come along in future, and hot-swap dependencies in and out.

And there's a pattern called 'dependency injection' that you can use to make your systems more composable / modular, so that hot-swapping can be configurable, rather than needing rebuilds etc. It's also used heavily to make software easier to test, too – you can swap 'stub' modules to simulate behaviour for integration testing purposes, for example.

So the 'plug-and-play' effect you're after can potentially be achieved using quite different pieces of software, as long as what they actually *do* is broadly similar.

There's been a whole bunch of work done on this over the years – take a look at the work of Martin Fowler in the design patterns space and you won't go too far wrong.

Adi Alter
on [20 July, 2017 at 14:47](#) said:

Thanks, Somaya, for the excellent summary. As the Rosetta Product Manager, I can assure you that the requirements you have listed are at the top of our list as well, and that many of these needs are already addressed in our system.

I fully agree with you that automation is a must here, and for this, integrations and standards are key. If I may add my two cents here, then I would like to emphasize the need to also standardize the preservation tools used by the different preservation systems. Currently there are multiple tools out there performing tasks such as format validation and identification, characterization, content migration, etc. Apart from the fact that some of these tools overlap in their functionality, each of these tools uses different technologies, is deployed differently and exposes a different API. If we could agree on standards here, on a single set of APIs that each such tool should expose, e.g. for a valid/invalid format, for extracted metadata, or for performing format migration, then deploying such new tools could be almost a plug-and-play process. This will save a great deal of time for both software vendors and system users, ensure that new tools will constantly be integrated, and enable users to focus on actual preservation work and not on system integration.

somaya
on [14 July, 2017 at 14:36](#) said:

Hi Ed,

Thanks for your comments. I'm neither an archivist or a record-keeper, so the skillsets of my DPOC colleagues Lee and Edith etc. are important. I have worked alongside archivists for many years, supporting the work they do.

Absolutely, authenticity cannot be sneezed at – and I've yet to find a suitable way of representing this wholly in the digital preservation space.

I do feel the digital preservation community is more au fait with fixity (which in itself does not prove authenticity in the slightest, but helps to build a picture of when authenticity of a file might have changed after being placed under control).

Perhaps this section on Provenance should really have been titled "capturing file system metadata" as authenticity and provenance go hand in hand. And, as you say, custodial history is also significant and needs retaining.

As for EDRMS', for those organisations lucky to have such a system, they are incredibly important. Both Lee and I have been vocal on this subject. We – as a field of practice – need to put some time into what should be pulled from EDRMS' and where this metadata is best placed (and where it can become the 'source of truth'), as some of it is the domain of an Archival Management System and some is a Preservation System. There are several differing views on this within the digital preservation community – and have been (and are likely to continue to be) the topic of some lively debates.

Ed Pinsent
on [14 July, 2017 at 13:56](#) said:

Very good piece of work Somaya...

I agree with you 100% about provenance, which is one of my preoccupations as an archivist...I also think what you describe in this section aligns with what I would call "custodial history".

I have spoken before about EDRM systems, and wondered out loud if an export from an EDRMS would give me a full report on every document stored in it, the original directory structure, folder and file names, audit trails, names of editors, revision histories, etc. If I can't get that, my content might lack a custodial history, and this starts to cast doubts on its provenance.

I would extend "provenance" to include the work done by a traditional archivist too, by which I mean an attempt to rebuild (in the form of an ISAD(G) catalogue) the administrative history and structure that led to the creation of the resource, along with an accurate description of its original order.

I should like these things – and other related aspects – to be increasingly taken into account in our thinking about digital preservation, and in preservation systems too. If you're saying there's a provenance gap, then we're in complete concord. It's all in the name of supporting authenticity, which I am certain you are interested in also (even if not explicitly stated here).

ehalvarsson
on [14 July, 2017 at 09:05](#) said:

An excellent summary Somaya! I very much enjoyed reading this.

somaya
on [13 July, 2017 at 17:20](#) said:

Hi Jon – Thanks for your response. There are many additional requirements I would add – but as I needed to keep this to a handful, these are the things that made it onto the shortlist.

In regards to your point, this ties into what I've said about about automation and prioritisation for Error Handling. We do need preservation systems to help us do our work more effectively.

That said, I think that automation of file format migration must take into account the type of object that the file is part of. Is it a stand-alone file or set of files all made up of a single format, or is it one part of a complex object that has dependencies and the potential for linkages to break? The example of this I've been dragging out of the cupboard for a decade or two is the Macromedia Director Projector with a series of linked videos... if the video files are migrated to another format (e.g. from MOV to AVI), then the videos won't playback and the the work is 'broken'. Of course, any website is a good illustration of this issue too.

So, the context of the digital materials is really crucial too. Having subject matter experts working with people doing acquisition, curation and preservation in the digital space is important.

As for the other things on my list, I could go on – but I'll leave it at that.

Jonathan Tilbury

on **13 July, 2017 at 15:21** said:

Somaya, this is a really good summary. As CTO of Preservica we either are on the way or working towards many of these goals.

I think the only one I would add is more automation of file format migration based on community defined best practice, so the factual information and best practice policy information are agreed by the community and then automatically applied by each system.

Regards

Jon

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