Parsimonious preservation: preventing pointless processes!

(The small simple steps that take digital preservation a long way forward)

Tim Gollins

Head of Digital Preservation, The National Archives, UK

Abstract

While there are many and varied threats to the successful curation of digital material, the impression given by the current generation of digital preservation systems and by much of the "received wisdom" in the digital preservation community is that imminent technological (software/data format) obsolescence is the primary threat. This gives rise to the belief that the only way to successfully start doing digital preservation is to invest in a technically complex, expensive, and difficult to operate integrated digital preservation system. This paper argues that, while the threat of technological obsolescence is real in some particular cases, a much more imminent threat is poor capture and inability to achieve safe and secure storage of the original material. By applying the principle of parsimony to digital preservation, institutions can find ways forward that are incremental, manageable and affordable, and which achieve the goal of securing our digital heritage for the next generation.

The paper observes that many existing institutional IT systems (and their support teams) provide as a part of normal business the capability to address many of the challenges of capture, custody, and integrity facing the new digital curator. The paper also argues that open source or free resources can be applied intelligently to further address these challenges, without needing huge integration or significant IT resource allocation. The paper gives an example of this and argues that the simplicity of a digital preservation system itself is critical to ensuring the long term access to the material that it holds. This approach makes it possible for even the smallest institution to begin to take steps to ensure the long-term survival of our vital digital heritage.

Why parsimonious?

Apart from obvious alliterative opportunities in the title, we choose to adopt the principle of parsimony (as first put forward in the 14th century by William of Occam) to guide our work on digital preservation. The word parsimony is defined as "economy in the use of means to an end; especially: economy of explanation in conformity with Occam's razor" and that implies not looking for solutions to problems for which evidence is absent, and using only the minimum necessary intervention to secure our digital heritage for the next generation. This is not a "miserly" or "stingy" approach as some definitions of parsimony would imply, however it does have the benefit of thrift in these challenging economic times (Merriam-Webster Online Dictionary, 2009).

To apply the principle of parsimony to digital preservation we first need to consider the scope, our goals, and the evidence for actual threats to their achievement. We should also remember that the principle of parsimony is just that, a principle, an heuristic, a rule of thumb to help us understand and manage our world, but not a rigid doctrine.

Scope and audience

This paper specifically focuses on the threats and challenges to the *preservation* of digital data (that is ensuring it remains available to future generations). The enabling of ubiquitous and universal online access to such collections is a separate matter and is not addressed here. Some will argue that there is little point in preservation without access; I would argue that there is little point in access without preservation. Given limited resources, preserving and keeping material available in a limited context should be a priority; this at least provides the potential to provide access in a wider context at a later date.

This paper is not intended to specifically address the concerns of the largest digital heritage institutions (national libraries and archives), nor those of large national or international institutions which create vast quantities of sophisticated or specialised data (such as large particle physics institutions or meteorological centres). These institutions are in classes of their own and their scale of operations bring further challenges; whilst the concept of parsimonious preservation should influence the design of their digital preservation systems the smaller scale approaches described here will not meet their needs.

This paper focuses on institutions that routinely preserve data created during the normal business of administering or governing a company, educational establishment, government organisation, or third sector institution. It also focuses on those smaller institutions faced with the challenge of starting out on the road of digital preservation, in particular those in the cultural heritage sector.

The goal — forever!

How long can an institution realistically plan to keep things for? It can set a long term aim; indeed its charter may require it to do so, but in practical terms how far ahead can it really plan? I contend that, while the overall aim may be (or in our case must be) for "permanent preservation", "in perpetuity" or "forever", the best we can do in our (or any)

generation is to take a stewardship role. This role focuses on ensuring the survival of material for the next generation — in the digital context the next generation of systems. Here immediately the principle of parsimony can be applied; the minimal intervention implied means minimal alteration, which brings the benefits of maximum integrity and authenticity. It also means a minimal assumption as to what the future may bring or enable; the one thing history teaches us is that predicting the future is really problematic! This is the same principle that is applied by us in the care of a physical collection of artefacts (e.g. paper documents) (The National Archives (a), 2009). We should also remember that in the digital context the next generation may only be five to ten years away!

Threats — immediate and real

There are many complex and interacting threats to the long term survival of digital objects. However these threats tend to boil down to a combination of the following (in no particular order):

- media (removable) decay/obsolescence;
- hardware obsolescence;
- software/data format obsolescence;
- online storage disaster/decay;
- incomplete/inadequate capture.

All of these threats are real to a degree. However not all of them are immediate and pressing for the majority of institutions or the majority of data, even with material that is relatively old (in digital terms). The most pressing and immediate threat to digital data is incomplete or inadequate capture:

- in other words "Don't it always seem to go that you don't know what you've got till it's gone?" (Mitchell, 1970);
- and as Bracton said in the 14th century "vulgariter dicitur, quod primo opportet cervum capere, et postea cum captus fuerit illum excoriare" or "it is commonly said that one must first catch the deer, and afterwards, when he has been caught, skin him";
- although it turns out Mrs Beaten never did say "First catch your hare"! (Answers.com, 2009).

This is so much a matter of common sense that it can be overlooked! We can only preserve and process what is captured. While this has always been the case, digital information brings with it opportunities that we should consider at length, before continuing to adopt our existing capture policies. Opportunities to analyse complete sets of data (as opposed to small samples), unencumbered by the difficulty of analysing or having to store enormous volumes of paper, can fundamentally change the value of a potential collection.

Next to consider is the decay or obsolescence of removable storage media. This is one of the most dangerous threats to digital data; it catches you unawares, and only manifests itself at the point when you can do very little about it! We all have them at home, the 3.5" floppy disk, the Zip Disk containing our dissertations, or at work the personal DVD back up we took only four years ago. It may already be too late! At least one well respected national archive has already experienced a tape media failure of this kind, and was not quite able to recover all of the data on the media even after very considerable expense using specialised digital forensic recovery contractors (Gollins, 2008).

Moving on to consider online storage disaster or decay (so called "bit rot"); although disaster is theoretically a significant threat (the consequences of an *unmitigated* online storage hardware failure would be catastrophic), online storage environments are almost always specifically managed to mitigate the risk of such failures (be that through use of RAID and/or offline back up regimes). Bit rot (where, as a result of random physical processes a bit of data is flipped from 0 to 1 or vice versa) is theoretically an issue, but only becomes of any statistical significance in the case of very large collections. In practical circumstances, for the majority of institutions, the measures already taken by a good IT services department will more than adequately mitigate these threats.

Hardware obsolescence, when not directly associated with some form of removable media, is also a much less pressing problem, and tends to manifest itself in relatively rare circumstances where specialised hardware is needed to display unusual forms of data. For mainstream data on mainstream systems I contend that it is not a significant issue (Rosenthal, 2009).

Threats – future or mythical?

Finally we come to software (or data format) obsolescence; this is perceived to be a very significant and imminent threat. It is my contention that this threat is significantly smaller in practice, for the majority of data in the majority of institutions, than the perception or received wisdom would indicate. This view is based on the experience of The National Archives over the last 10 or so years, and the experience we are beginning to get as we scale up our ability to accession new born digital records into the archive.

And it is not just our view; at this summer's SUN PASIG (Preservation Special Interest Group) meeting in Malta (Sun PASIG, 2009) David Rosenthal of Stanford University compared the stability of the UNIX File system interface with the vision of obsolescence envisioned by Jeff Rothenberg in 1995. Jeff's vision (Rothenberg, 1995) was that "... digital documents are evolving so rapidly that shifts in the forms of documents must inevitably arise. New forms do not necessarily subsume their predecessors or provide compatibility with previous formats." Rosenthal characterised this as a view that "incompatibility is inevitable, a force of nature". In challenging this view Rosenthal observed the longevity of the UNIX file system. With a defined interface now some 30 years old, capable of handling disks 1,000,000 times bigger than when first created, and executed by new software at least four times bigger (but faster and more reliable) than the original, it is still capable of reading every single disk ever written in that 30 years (Rosenthal, 2009). Rosenthal also observed that the open source movement, online publication, online storage, and other similar developments, all strongly mitigate the imagined format obsolescence risk. Looking back with 20/20 hindsight at Rothenberg's paper Rosenthal concluded "format obsolescence almost never happens".

New digital curator's response

So what should be our response to these threats? What might a parsimonious preservation system look like in a small or medium-sized institution? I believe that this is where the benefits of applying the principle of parsimony really begin to show.

Firstly consider our goal; not "to preserve forever", but the more parsimonious "to ensure the survival of digital material for the next generation" (of IT systems). Since many

institutional IT systems (and preservation systems are no different) last for between seven to ten years (often with a conservative hardware upgrade at the 4/5 year point) this begins to look much less scary.

From the stewardship perspective, one of the critical capabilities to develop is the ability for data to be extracted at the end of life of a proposed system in a complete form, in bulk, and without the need to modify or transform it. This is the final act of a good and responsible steward; to pass on the objects in his or her care to the next steward.

A thrifty response to the preservation challenge would be to ask "what have we already got that will do the job?"; I would respond by asking you to examine an institution's generic IT/desktop infrastructure and the organisation and systems that support it. You might be surprised how far this will go in meeting the challenges.

Taking the storage threats first; most networked desktop systems that run across an institution will have some volume of corporate storage, usually quite a considerable amount. This storage is ideal as the basis for your preservation system. It will undoubtedly be reliable, backed up, and migrated to new live media when required, by a team of competent professional IT support staff. It will have access controls to prevent unauthorised modification (maintaining integrity and authenticity), and a process for allocating this control. What it may not have is space; however it will usually have someone who is responsible for sizing and capacity, who will know how to get more and who will have a process to do so. This may of course cost some money, but although we are being thrifty we may not be able to do all of this on a zero budget.

Now, having made these observations about an institution's current infrastructure, let's look at the threats from obsolescence. We have established that the hardware is most unlikely to be a problem since it is modern and updated in a managed way, and such changes rarely introduce issues. What about removable media? We know that this is the most threatening form of obsolescence, but the answer is simple. Do not store removable media at all; copy the data onto the networked storage. There are some issues to be addressed, but these are more than manageable with good capture practices; "what about the authenticity of the object received?" I hear the traditional archivists cry. Finally, then, in the obsolescence area we come to software based or data format obsolescence. I believe that any residual risk in this area can be effectively managed by good information management and good capture practice at the beginning.

So let us consider capturing a collection of digital information; what are the issues we should address? This can be summarised by addressing the task of "knowing what you have got". Aspects of this include:

- Knowing what the subject matter of the collection is (including if there are any significant coherent sub-collections).
- Knowing how complete is the collection (is it a sample of a larger whole, and if so on what basis).
- Knowing the context for the collection (where it came from, time context, and perhaps some political, social and economic context of the source).
- Having a complete inventory of the collection, in particular:
  - a list of all the file names and any "structure" (e.g. a record of any directory or folder structure and the file's position in that structure);
  - the data format of each file;
  - the date of each file's creation and last modification;
  - a "signature" or checksum of a reputable kind, to "seal" each file and demonstrate its integrity.
- Knowing that there are no viruses or other "infections".

This represents a starting point; clearly if more detail is available regarding the subject matter of the collection at a more fine grained level (individual subject metadata for each file or document), then this should be recorded and kept along with the collection. All of the above are achievable for really quite significant volumes of data, using simple "record keeping", some free software, and well documented procedures. What free software exists out there to help with this? Most of the challenges of an inventory can be met using DROID (The National Archives (b), 2009).

DROID was originally produced by The National Archives to enable the unambiguous identification of the format of a data file. It is free, open software, released under a BSD licence. At the time of writing a consultation is completing, to determine the features to be introduced in the next release of DROID (version 5) (The National Archives (c), 2009). DROID can already produce a profile of a shared drive or file system, listing the files, their identified type and other simple metadata (creation date etc.), and one of the proposals for DROID 5 is to incorporate Digital checksums or hashing to enable integrity of data to be checked. DROID uses information drawn from The National Archives' PRONOM (The National Archives (d), 2009) database, to enable the identification of a wide variety of file formats. Work continues at The National Archives to increase the number of formats covered through our own efforts and through partnering with other well respected institutions wherever possible.

If DROID is not your choice, then there are a number of other free checksum and digest creation tools that a trivial internet search will reveal (I have not listed any specifically as I cannot vouch for the quality of any particular product, as we do not use them). Any of the well known virus check software is appropriate to use to ensure that proposed content is virus free; however a "quarantine protocol", where received material is held safely for a month or so is strongly advisable. This enables any novel virus (that might have infected the content) to have been characterised and signatures to have been received by the virus checker before acceptance of the content into the institution.

A further important "parsimonious" consideration is what types of information (or formats) the institution should accept. I believe that an institution should only accept material that at the point of acceptance it already possesses the ability to read. This is not a point about software obsolescence — the data is probably still completely accessible using the right current technology — it is, however, a point about the complexity and variety of software any one institution is capable of maintaining and supporting at reasonable cost. It is also a point about the age of removable media (and its obsolescence and decay) that might have been used by the source (or donator) to store the information before presenting it to the curating institution. The costs of digital archaeology (or digital forensics) can be extreme, and value for money must be a pragmatic consideration for any institution.

Taking all of the above into account, I would contend that the barriers to effective entry into the activity of digital preservation are much lower that might at first be believed, and that in short very many institutions are in a position do digital preservation now.

Added benefits of parsimony

What other benefits accrue from a parsimonious approach to preservation? An important strategic consideration when considering a digital preservation system is the potential costs of maintaining the preservation system itself. Complex systems, while they may possess many useful functions, are by their very nature more costly to maintain. The parsimonious vision presented here implies simple systems composed of ubiquitous components, which are held together by robust and pragmatic business processes. Such systems are often very cheap to maintain (by comparison), and may well be maintained as a part of "business as usual" by an institution's own IT department.

A further benefit of a parsimonious approach to preservation is the minimal degree of intervention it implies with respect to the individual information objects themselves. From an information theory perspective, any modification to the bit sequence of an information object potentially destroys information. There is an argument with respect to the significance of different elements of information in the object, however this is always a qualitative argument, and if preservation can be achieved by doing nothing to the object then I believe this must be preferable. This approach is also likely to result in a collection of information that is inherently able to address the critical stewardship issue of passing the collection on to the next generation.

Conclusion

I have argued that the imminent threats in digital curation for institutions new to the field are other than they might first appear; in particular while the threat of technological (software/data format) obsolescence is real in some particular cases, a much more imminent threat is poor capture and storage of the original material in a safe and secure way.

I have observed that the capabilities that many existing institutional IT systems (and their support teams) provide as a part of normal business often address many of the challenges of capture, custody, and integrity facing the new digital curator. I have given an example of how open source or free resources can be intelligently applied to further address these challenges without needing huge integration or significant IT resource allocation.

In short, a series of small, simple and affordable steps can be taken by institutions to ensure the long-term survival of vital digital data, thus lowering the barrier to entry for institutions to the interesting and vital aspect of information management.

References

Answers.com. 2009. First catch your hare. *Answers.com.* [Online] 2009. [Cited: 7 September 2009.] http://www.answers.com/topic/first-catch-your-hare.

Gollins, Tim. 2008. Tape media failure. *private briefing recieved from a european national archive.* October 2008.

Merriam-Webster Online Dictionary. 2009. parsimony. *Merriam-Webster Online Dictionary.* [Online] 2009. [Cited: 7 September 2009.] http://mw1.meriam-webster.com/dictionary/parsimony.

Mitchell, Joni. 1970. Big Yellow Taxi. *Roots of Bob Dylan.* [Online] 1970. [Cited: 7 September 2009.] http://www.bobdylanroots.com/bigyellow.html.

Rosenthal, David. 2009. Presentations from the PASIG Summer Meetings in Malta. *Stanford Universtity Libraries and Academic Information Resources.* [Online] 24-26 June 2009. [Cited: 8 September 2009.] http://lib.stanford.edu/files/rosenthal_pasig_lockss.pdf.

Rothenberg, Jeff. 1995. Ensuring the Longevity of Digital Documents. *Scientific Americian.* Janurary 1995, Vol. 272, 1.

Sun PASIG. 2009. Sun PASIG (Preservation Special Interest Group). St Juliens, Malta : Sun Microsystems, 24-26 June 2009.

The National Archives (a). 2009. The National Archives Preservation Policy. *The National Archives.* [Online] 12 June 2009. [Cited: 7 September 2009.] http://www.nationalarchives.gov.uk/documents/tna-corporate-preservation-policy-2009-website-version.pdf.

The National Archives (b). 2009. DROID sourceforge project page. *Sourceforge.* [Online] 2009. [Cited: 16 September 2009.] http://droid.sourceforge.net/.

The National Archives (c). 2009. DROID 5.0. *DROID 5.0 Wiki.* [Online] 2009. [Cited: 16 September 2009.] http://droid5.yourwiki.net/wiki/DROID_5.0.

The National Archives (d). 2009. PRONOM Technical Registry. *The National Archives.* [Online] 2009. [Cited: 16 September 2009.] http://www.nationalarchives.gov.uk/PRONOM/Default.aspx.

Contact

Tim Gollins
timothy.gollins@nationalarchives.gsi.gov.uk