ARIADNE

Excuse me … Some Digital Preservation Fallacies?

Chris Rusbridge argues with himself about some of the assumptions behind digital preservation thinking

Introduction

Excuse me...

I have been asked to write an article for the tenth anniversary of *Ariadne*, a venture that I have enjoyed, off and on, since its inception in 1996 as part of the eLib Programme, of which I was then Programme Director.

Some years ago I wrote an article entitled "After eLib" [1] for *Ariadne*. The original suggestion was for a follow-up "even more after eLib"; however, I now work for JISC, and that probably makes it hard to be objective!

In "After eLib", I wrote this paragraph about digital preservation:

> "Back to the Electronic Libraries Programme, what were some of my favourite projects (I won't say best; this is definitely a subjective list)? The project of greatest personal satisfaction for me is CEDARS [2], the digital preservation project. Ensuring the long-term existence of digital materials was not an element of the Follett report, and this seemed a significant gap when I started thinking about applying for the job of Programme Director. Others were also aware of the importance of this area, most particularly FIGIT's Chair, Lynne Brindley, now CEO of the British Library. We still have CEDARS as an exploratory project piloting ideas (which together with its JISC/NSF-funded companion, CAMiLEON [3] has a high international profile) rather than a full-blown digital preservation service; this is another example of the difficulty of taking even widely supported ideas through research into service. It is also true that the technical problems in this area are not yet solved, but also that the real problems are organisational and political rather than technical. I hope that in this interim period our consciousness of the problem is raised enough that temporary expedients will be found so that little of importance will be lost."

I now work for the Digital Curation Centre, which is concerned to help improve support for digital preservation and curation. In my first year, I have had occasion to listen to many presentations on aspects of preservation in particular, and to read many articles and other texts. In the process, I had occasion to look for the outcomes of an eLib Project, Infobike. The eLib Programme pages still exist [4], and the description of the project in those pages also exists [5]. However, the project web site refers to does not exist. The UK Web Archiving Consortium, which is sponsored by JISC and includes some eLib projects, has not been able to archive the outcomes of this project. We have to go to the Internet Archive's Wayback Machine to find archives of the web

site for the Infobike project, dating from January 1998 to August 2003 [6]. There I was able to find a general description of the project, an architecture diagram, and a description of the system components. Nothing fancy was needed; just access to the resource, and a current web browser. This re-awoke the train of thought identified at the end of the "After eLib" paragraph above: what is the use of all this grand digital preservation theory, if we lose access to the data itself?

Since then, a number of common assertions, or perhaps assumptions, about digital preservation have begun to worry me. No one person has said all these things, but increasingly they seem to be in the background of conversations. I will put these forwards as a list of statements, but, in some respects at least, I think they are fallacies:

1. Digital preservation is very expensive [because]
2. File formats become obsolete very rapidly [which means that]
3. Interventions must occur frequently, ensuring that continuing costs remain high.
4. Digital preservation repositories should have very long timescale aspirations,
5. 'Internet-age' expectations are such that the preserved object must be easily and instantly accessible in the format de jour, and
6. The preserved object must be faithful to the original in all respects.

These statements seem reasonable, and perhaps they are. However, I feel we might benefit from a rather jaundiced look at them. So that is what I thought I would attempt for this article. Beware, the arguments presented here are not settled in my mind; indeed this is to some extent part of an argument with myself!

Digital preservation is very expensive? Excuse me…

Is digital preservation expensive? It depends whether you compare it with print preservation! Two thought experiments are worth doing at this point. First, think about your nearest and dearest research library, national library, or research archive. It will be a big building (a very big building, often rather grand), often with a comparatively small proportion of space inside for people. Think about the number of librarians, archivists etc who look after the stock in those buildings. Just calculate how much it must cost! David Thomas of The National Archives [7] has written: "Storing and conserving our physical stock of records (which has now grown to 176 kilometres) cost £14.3 million in 2002 […]. Retrieving a paper record for use by a reader costs about £6; delivering one over the Internet cost 13p" [8]. Yes, this is access and not preservation, but in the physical world these two are strongly bound together. The cost of the Atlas Petabyte Data Store [9] is a few million pounds; any major research library costs the odd hundred million pounds.

In the second thought experiment, imagine a digital world in which this wonderful new technology called the Basic Object for Organising Knowledge had recently been invented. You are head of information services for a major research university (providing all those services digitally, from the access services fronting your digital repository), and have to persuade your Vice-Chancellor to invest in a new facility for these BOOKs; maybe a couple of million of them (only a fraction of the numbers of objects in your digital stores). You can probably script the interview yourself… "You want a special building with 10 floors, with huge floor loadings and a special environment? You want 200 staff? You want how many million pounds? And after all

that, the users have to go **into** the facility to access these BOOKs? You must be kidding me; get out of my sight!"

My point is that all preservation is expensive, but we are used to it and accept it as part of the cost of a cultured and educated life… at least in the print world. The amounts of money being spent on preservation of digital material are comparatively tiny, and in any per-unit terms, will probably remain so. The trouble is, it is a new cost, and we have not worked out how to factor it into our budgeting and business models. My guess is that in the long term, we will realise that print preservation is very expensive, while digital preservation is comparatively cheap!

File formats become obsolete very rapidly? Excuse me…

There's a lot of rather panicky talk about the rapid obsolescence of file formats. Some of this is true, some is perhaps less so. To some extent it depends on your timescale (see 4 in the list of fallacies above).

I think we need to analyse rather carefully what we mean by file formats. On a simple analysis, I can find the following somewhat distinct cases (there are probably more):

1. Media formats
2. File formats created from hardware devices (eg digital cameras, scanners etc) and telemetry
3. File formats created by programmers for specific projects
4. File formats from standards-based, community or open source projects (perhaps not completely distinguishable from the previous case)
5. File formats resulting from consumer-oriented commercial software products
6. File formats from highly configurable products (SPSS is the example I have in mind)
7. File formats protected by Digital Rights Management systems, or other forms of encryption or proprietary encoding.

The list is long enough to appreciate that there are subtleties here. It is clear that the first three cases do provide significant risk of early obsolescence, and the last case certainly also represents significant risk of loss. There are significant risks in case 6 if the file is not looked after together with its attendant metadata or documentation (or if this never existed).

However, I think most people would assume that the dictum about formats becoming obsolete very rapidly applies particularly to case 5, file formats from consumer-oriented commercial products. I used to think this, too. But I have asked around, and I cannot find any good example of this class of file where the content is completely inaccessible today! So with this article I am inviting anyone with details of good examples of case 5, to respond to a posting on this topic on the Digital Curation Centre Associates Network Forum [10], or to email me at the address shown at the foot of this article.

Perhaps two things are happening here: one about why this fallacy is so widely held, and one about why it is perhaps less true than it might have been. Many of those concerned with preservation are (as it were) older people, who grew up in the pre-Internet period. Things certainly did change rapidly then. Managers were faced with making technology choices that did indeed appear to bind them into technological dead ends. There were many different options for everything, and interoperability was

rather weak. Change was rapid, as company after company went out of business or was bought up. Obsolescence was a real worry.

It seems to me that nowadays the move to a consumer market and the rise of mass access to the Internet have comparatively stabilised things. Somehow the system has gained a significant momentum that it did not have before. Cost of entry to markets has greatly increased, and choice and variety have decreased. The pace of new releases of mass-market consumer products has decreased. This may not be so true for all market segments (for example, this article was planned using one of the software products for mind mapping [11], a market segment where file format incompatibilities between products and even versions are rife), but it is increasingly true for those mass products which create most files of interest for preservation.

Note that my argument is about total loss of information content. There are clear examples where recovery of information from old files is partial or incomplete; see the Representation and Rendering Project Report by Paul Wheatley et al, for example [12]. It is possible that, with a concerted communal effort, we can do much better with some of these file formats; for example, the wide range of graphics formats now accessible in part due to the combined efforts of many individual enthusiasts.

Part of the key here is to collect and to share information. This is where the several efforts to gather representation information in registries are so valuable. The first such major effort was PRONOM [13], from The National Archives in the UK; in the near future the Representation Information Registry and Repository from the Digital Curation Centre [14] is expected to come on stream, and we have recently heard that the Global Digital Format Registry Project [15] from Harvard University Libraries and others has been funded by the Mellon Foundation. If these registries can find ways of sharing information, and of dividing up the problem space without remaining reliant on one another, we should be able to make good progress.

There may well be two flaws in my argument: genuinely disruptive technological change, and extended time. The Internet and the consumer mass market that emerged in the early 1990s could scarcely be imagined before, and have had radical effects on the way things work. We should expect some such change to arrive in the next 10 or 20 years, and throw any of our cosy predictions (and plans) off track. And clearly, if enough time passes, then these problems of inaccessible formats will emerge in one form or another. However, time is an issue for preservation repositories in many ways, and is the subject of the next section but one.

Interventions must occur frequently? Excuse me…

This fallacy follows from the last. Simply, if file formats become obsolete and inaccessible rapidly, then digital preservation interventions to reduce loss must occur frequently. The KB of the Netherlands [16] has suggested in its cost model that file migrations might be needed as often as every 3-5 years. However if, as argued above, the whole system is gaining sufficient inertia to stabilise partially, then it is a reasonable bet that file formats current today, if chosen with a little care, will still be accessible in 10 to 15 years time. This certainly seems to be the case now; although most people with access to older files (say 10 years or so) can cite cases of some difficulty in accessing the content of some of them (for example Microsoft Office version 4 file formats), these are generally not insurmountable.

Some may see this as a highly dangerous argument, encouraging complacency. There is certainly a risk (and complacency itself would be a very high risk strategy!), but the arguments about the continuing high cost of digital preservation are also a serious deterrent, to which the answer too often is to throw up one's hands and say "Can't be done!".

Investment in digital preservation is important for cultural, scientific, government and commercial bodies. Investments are justified by balancing cost against risk; they are about taking bets on the future. The priorities in those bets should be: first, to make sure that important digital objects are retained with integrity, second to ensure that there is adequate metadata to know what these objects are, and how they must be accessed, and only third to undertake digital preservation interventions. This does tie in with my final fallacy, raising the question of the extent to which the costs should be loaded onto the archive or the end user. However, first it is worth thinking a little more about timescales.

Digital preservation repositories should have very long timescale aspirations? Excuse me…

Much of the literature on digital preservation assumes very long time scales, sometimes of hundreds or even thousands of years. One sees comments that suggest one of the possible risks a repository must guard against is the loss of the English language, for example; so it can be suggested that part of the representation information to deal with such cases would be an English dictionary (it is amusing that they still sometimes pre-suppose an Internet and Web).

In practice, until very recently almost all digital preservation was funded on short-term project money. David Giaretta, Associate Director for Development of the Digital Curation Centre has wisely remarked that the primary resource needed for digital preservation is **money** [17]. In practice, the largest risk to digital preservation is indeed money. Who has the resources to make a hundred-year digital preservation promise? Who can make an investment case with a hundred-year return?

The money problem has another side-effect. The more money that needs to go into expensive infrastructure, the less is available for addressing the real risks to digital objects. Rosenthal et al point out "Few if any institutions have an adequate budget for digital preservation; they must practice some form of economic triage. They will preserve less content than they should, or take greater risks with it, to meet the budget constraints. Reduced costs of acquiring and operating the system flow directly into some combination of more content being preserved or lower risk to the preserved content." [18]. So designing for very long timescales itself has the potential to cause loss.

It is true that we are beginning to see the emergence of digital preservation repositories that can properly argue they have a hundred-year timescale. Who could doubt that the British Library, The National Archives, and other national memory institutions have long-term intentions? But even they are not immune to the effects of disruptive technology.

Another thought experiment may be helpful here, for those of you old enough. Cast your mind back to the early 1990s. This was the period immediately prior to the Internet, when gopher was king, and the World Wide Web appeared no more likely to be a successful technology than WAIS (Wide Area Information System). Who at that

time could have imagined the world of today? Who, planning a hundred-year digital preservation repository in 1992 would have made decisions we would think correct today? What makes us think we could do any better now? There is a strong tendency to project the current situation forward (and it might be argued I have done just that, above).

It seems to me that it makes more sense for most of us to view digital preservation as a series of holding positions, or perhaps as a relay. Make your dispositions on the basis of the timescale you can foresee and for which you have funding. Preserve your objects to the best of your ability, and hand them on to your successor in good order at the end of your lap of the relay. In good order here means that the digital objects are intact, and that you have sufficient metadata and documentation to be able to demonstrate authenticity, provenance, and to give future users a good chance to access or use those digital objects.

The preserved object must be faithful to the original in all respects? Excuse me…

One of the key ideas of the CEDARS Project [2] was "significant properties"; another (from OAIS, the Open Archival Information System [21]) is the "designated community". Digital objects (viewed as data structure plus mediating software) have a huge number of possible behaviours. Think of all the capabilities of a word processor such as Microsoft Word, operating on a digital document. During the creation phase of the document, a subset (probably not a huge subset) of those capabilities is brought into play. Other capabilities remain unused, but as long as the file remains in an environment where it can be accessed with the same software, those capabilities can potentially be used. Some of those capabilities, such as extracting a change history, may be important for some potential users. Other users may only want the capability to read the document, or perhaps to cut and paste extracts into other documents (an even smaller subset of capabilities than the creator needed).

The problem here is that there is no way of precisely defining the designated community, and similarly no way of foretelling the properties that future users might deem significant. This leads to pressure for preservation that must be faithful to the original in all respects.

Similarly, the Internet paradigm of instantly clickable, accessible results also seems to be applied as a "must have" aspect of preservation. The combination of full capability of digital objects preserved from the past, instantly available in today's environment, may be an ultimate goal, but is extremely expensive. As already noted, high-cost preservation means fewer digital objects preserved.

This situation has its resonances in the print world as well. Take a book such as Sir Walter Scott's *Kenilworth*, for example [19]. Scott was keen to publish anonymously, and so each chapter was sent to a different copyist, to disguise Scott's hand-writing; these chapters were then sent by the copyists to the printer, typeset and assembled. The resulting book, set in the heavy type-faces of the time, bound in leather, and full of errors, would be a daunting read for most of us, but of huge interest to the Scott scholar. Happily for the less scholarly, modern editions are widely available; they identify Scott as the author, and aim to "correct" many of the errors in early editions. So, the appearance, weight, pagination, authorship, publisher and text of the modern edition differ substantially from the original publication. Despite these changes, we

are content that this very different artefact represents the same "work" as the original. For us, the *story* is the significant property. For the scholar, the original is essential. The Scott scholar and the general public are, in this case, quite separate and distinct designated communities.

In the print world, these designated communities are served over the long term by very different kinds of preservation activity. The general public can be well served by the "preservation by diaspora" of the international library system. Lots of copies of books, perhaps in multiple editions, do indeed keep the significant property of the work safe. The scholar needs access to the few remaining copies of the early editions; preserved in Special Collections, in expensive controlled environments, accessed perhaps in special reading rooms supported on acid-free mounts, handled with special gloves…

It is true that the modern edition in my print example required the work of the scholar on the early editions. However, that scholar had to be prepared to do much more than the general public to access those editions, including perhaps travel to several libraries to study different copies. In other cases that scholar might have been required to learn ancient languages, or to decipher faded documents in archaic handwriting. Scholarship is a serious activity; it is potentially difficult. PhDs are awarded for contributions to scholarship, 3 or more years of painstaking research.

Back in the digital world, our scholar might be desperate to retain the functionality to extract information like change history from the digital object, but our general reader might be quite content with a much smaller set of properties. John Kunze and colleagues at the California Digital Library have suggested the idea of "desiccated" formats [20], versions of digital objects with much reduced sets of significant properties, but which are much easier to preserve.

There will be some repositories that rightfully aspire to preserve full functionality for many important digital objects. However for many repositories this will be too expensive a proposition. For them, as suggested above, the right approach (the right 'bet') may be to keep the original data files, the authentic original bit stream. When technology moves to a point where maintaining the capability to access these files is a problem, decisions on significant properties may mean that desiccated format copes should be produced. Bearing in mind the digital preservation mantra: "always keep the original bits", those requiring significant properties not in the desiccated versions have the possibility of investing to extract that information by performing their own transformations on those original bits, guided by the metadata and documentation available.

In the long run, as with Special Collections and archives, it is likely that the majority of preserved objects are very little (or perhaps never) used. Maintaining these objects in an instant readiness state is money wasted. Given the cost pressure on digital preservation, the aim should be to minimise the ongoing cost, to make easily preservable, desiccated versions if interventions are needed, and to put the cost of wider ranges of significant properties onto the user who demands them.

Excuse me…

Some of these remarks may be felt by colleagues to be almost heretical, and possibly damaging to The Cause. Excuse me, if that is the case, but given their importance and implications, I believe these issues still need analysis (almost certainly more careful

analysis than is expressed here). So, after these ruminations, how could I re-state my original set of possible fallacies? How about this?

1. Digital preservation is comparatively inexpensive, compared to preservation in the print world,
2. File formats become obsolete rather more slowly than we thought
3. Interventions can occur rather infrequently, ensuring that continuing costs remain containable.
4. Digital preservation repositories should have timescale aspirations adjusted to their funding and business case, but should be prepared for their succession,
5. "Internet-age" expectations cannot be met by most digital repositories; and,
6. Only desiccated versions of the preserved object need be easily and instantly accessible in the format de jour, although the original bit-stream and good preservation metadata or documentation should be available for those who wish to invest in extracting extra information or capability.

The key message that I want to get across in this article is that lack of money is perhaps the biggest obstacle to effective digital preservation. Assumptions that make digital preservation more expensive reduce the likelihood of it happening at all. Poor decisions on how investment is applied can have major implications on how much information can be preserved, and how effectively. Sometimes the right choice will be "fewer and better", as in Special Collections, for national memory institutions and major research libraries. Sometimes the right choice will be "cheaper and more". Repositories do have a choice, and must consciously exercise it.

References

1. After eLib, Chris Rusbridge, December 2000, *Ariadne* issue 26, http://www.ariadne.ac.uk/issue26/chris/
2. CEDARS Project http://www.leeds.ac.uk/cedars/
3. CAMiLEON Project http://www.si.umich.edu/CAMILEON/
4. eLib: The Electronic Libraries Programme: http://www.ukoln.ac.uk/services/elib/
5. The Infobike project description, eLib web pages: http://www.ukoln.ac.uk/services/elib/projects/infobike/
6. Wayback machine, Infobike project web pages: http://web.archive.org/web/*/http://www.bids.ac.uk/elib/infobike/homepage.html
7. The National Archives (TNA) http://www.nationalarchives.gov.uk
8. Digital Preservation at the National Archives, David Thomas, http://www.nationalarchives.gov.uk/preservation/digitalarchive/pdf/dpattna.pdf
9. Atlas Petabyte Data Store http://www.e-science.clrc.ac.uk/web/projects/petabyte
10. DCC Associates Network Forum posting http://forum.dcc.ac.uk/viewtopic.php?t=147
11. I first came across Mind Mapping through *Use Your Head*, by Tony Buzan 1974, British Broadcasting Corporation. See also http://www.buzanworld.org/mindmaps.asp
12. Survey and assessment of sources of information on file formats and software documentation; Final Report of The Representation and Rendering Project, Paul

Wheatley et al,
http://www.jisc.ac.uk/uploaded_documents/FileFormatsreport.pdf
13. The National Archives: PRONOM
http://www.nationalarchives.gov.uk/aboutapps/pronom/default.htm
14. Digital Curation Centre Development Representation/Registry Demonstration
http://dev.dcc.ac.uk/dccrrt/
15. Global Digital Format Registry (GDFR), Stephen Abrams and Dale Flecker,
http://hul.harvard.edu/gdfr/
16. *The KB Experience,* Erik Oltmans, Head Acquisitions & Processing Division, National Library of the Netherlands: presentation to the DCC Cost Models workshop,
http://www.dcc.ac.uk/docs/Eric_Oltmans.ppt
17. The Digital Curation Centre: Developing Support for digital curation, David Giaretta, http://www.dcc.ac.uk/docs/DCC-Development-Niees.ppt
18. Requirements for Digital Preservation Systems: A Bottom-Up Approach, David H Rosenthal, Thomas Robertson, Tom Lipkis, Vicky Reich, Seth Morabito, D-Lib Magazine, November 2005,
http://www.dlib.org/dlib/november05/rosenthal/11rosenthal.html
19. Kenilworth : a romance / By the author of "Waverley," "Ivanhoe," &c. In three volumes, Anonymous (by Sir Walter Scott), Edinburgh : Printed for Archibald Constable and Co.; and John Ballantyne, Edinburgh; and Hurst, Robinson, and Co., London., 1821
20. Overview of the UC Libraries Digital Preservation Repository (DPR). Kunze et al, California Digital Library,
http://www.cdlib.org/inside/projects/preservation/dpr/DPRoverview.pdf
21. Reference Model for an Open Archival Information System (OAIS), Consultative Committee for Space Data Systems, January 2002,
http://public.ccsds.org/publications/archive/650x0b1.pdf

Author Details

Chris Rusbridge
Director
Digital Curation Centre, University of Edinburgh

Email: C.Rusbridge@ed.ac.uk
Web site: http://www.dcc.ac.uk/