Putting Parsimonious Preservation into Practice

Recent Developments in Digital Preservation at The National Archives (UK)

Tim Gollins

Head of Digital Preservation, The National Archives, UK

Abstract

The principle of Parsimonious Preservation was developed at The National Archives (1) and now underpins advice and guidance given to the UK archive sector on digital preservation (2). The National Archives has recently applied this principle to the architecture and design of its next generation digital preservation environment, the Digital Records Infrastructure (DRI).

This paper re-caps the principle and offers additional evidence in support of Parsimonious Preservation in larger institutions and archives. It describes some of the challenges now facing The National Archives in the field of digital preservation and how the archive has responded.

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 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.

Parsimonious Preservation challenges the assumptions of traditional digital preservation and offers a sustainable, realistic response, for large and small institutions to the doom laden views that pervade the more traditional literature in the field.

The overriding message from Parsimonious Preservation is “Don’t Panic!”.

Introduction

Apart from the 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 (3).

To apply the principle of parsimony to digital preservation we need to consider our 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

The Principle of Parsimonious Preservation

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) (4).

We should also remember that in the digital context the next generation may only be 5 to 10 years away!

Threats – 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).

First let us consider 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 4 years ago. It may already be too late! Media may not last as long as the manufacturers claim (5) and even if it does the devices to read it may no longer be available.

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 storage hardware failure would be catastrophic), online storage environments are almost always 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 only an issue in the case of very large collections. In most practical circumstances, for smaller institutions, the measures already taken by a good IT services department will more than adequately mitigate these threats.

We should now mention hardware obsolescence; when not directly associated with some form of removable media, is also a much less pressing problem, and tends to manifest 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 (6).

Threats – Immediate

However, 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?*" (7)
- 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*"! (8).

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 even in the context of paper records, digital information brings with it opportunities that should be considered carefully, before blindly adopting existing capture policies. The National Archives have recently consulted on a new Record Collection Policy (9) with exactly this in mind.

It is important to note that even with a good collection policy in place, practical considerations of getting data from the organisation that created it with sufficient contextual information can present significant issues; this is the vexed and controversial topic of "metadata".

Threats – Distant

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, the experience we are beginning to get as we scale up our ability to accession new born digital records into the archive and more recently a survey of the records that we, as an administrative government department, create ourselves (see below).

And it is not just our view; at the SUN Preservation Special Interest Group meeting in Malta (10) in 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 (11) 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 4 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 (6). Looking back with 20/20 hindsight at Rothenberg's paper Rosenthal concluded "Format obsolescence almost never happens".

More recently in a number of Blog articles Rosenthal has re-confirmed his views and updated his arguments (12) (13) (14), amongst other sources he draws on the carefully framed and executed research from Andy Jackson at the British Library concerning the distribution of formats over time in the UK web domain (15).

Research on The National Archives Own Records

Recently, colleagues in the Knowledge and Information management Team at The National Archives conducted a survey of the file formats (as file extensions) of all of the records held in our own EDRMS. To be clear these are the records that we produce as a government department in our work developing public policy, guidance, and delivering a public service. We believe them to be typical of a UK government department; this is not a survey of our "collection".

The survey reveals the extremely well known "long tail" distribution of file extensions that conforms to power law distribution. Over 80% of the data is contained in the top three extensions, and over 99% in the top 30 (see Figure 1and Table 1); the tail continues for 800 extensions in total.

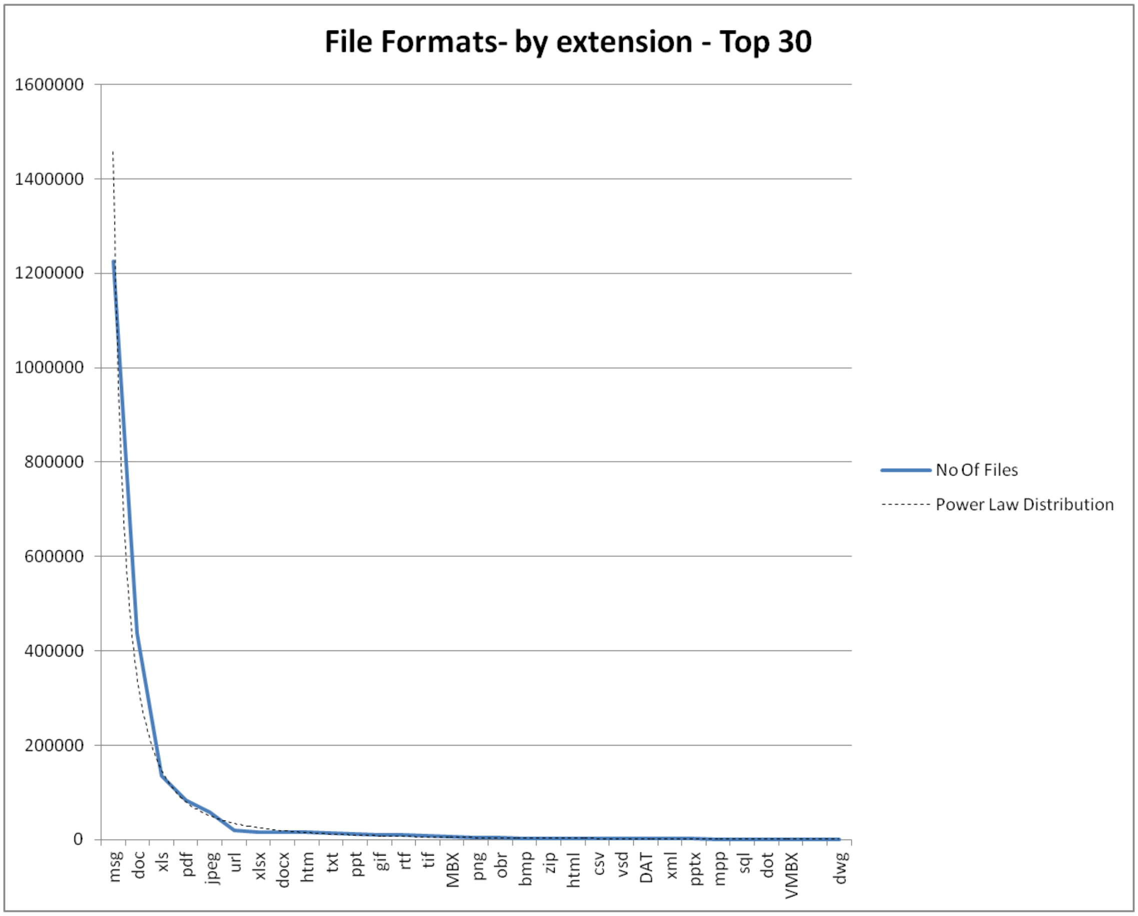

Figure 1

Extension	**Count**	**Type**	**Cumulative %**
msg	1225790	Email	57.6%
doc	437803	Word Processor	78.1%
xls	135573	Spreadsheet	84.5%
pdf	82524	PDF	88.4%
jpeg	58748	Image	91.1%
url	20701	Link	92.1%
xlsx	16777	Spreadsheet	92.9%
docx	16765	Word Processor	93.7%
htm	15616	Web Page	94.4%
txt	14199	Plain Text	95.1%
ppt	12834	Presentation	95.7%
gif	11154	Image	96.2%
rtf	11046	Word Processor	96.7%
tif	9293	Image	97.2%
MBX	6044	Email	97.4%
png	4812	Image	97.7%
obr	4368	Link	97.9%
bmp	3286	Image	98.0%
zip	2499	Zip	98.1%
html	2307	Web Page	98.2%
csv	2266	Structured Text	98.4%
vsd	2088	Visio	98.5%
DAT	1972	Data	98.5%
xml	1898	Structured Text	98.6%
pptx	1700	Presentation	98.7%
mpp	1557	Project	98.8%
sql	1417	SQL (database query)	98.9%
dot	1297	Word Processor Template	98.9%
VMBX	1263	Email	99.0%
	1134	(Blank)	99.0%
dwg	1072	AutoCAD	99.1%

Table 1

Examining the list of these top 30 formats we see things with which we are all familiar. We also observe that a number of these formats have been in existence for a long time (and many of the newer formats come from the same well respected sources). We further observe that these are formats that are common across the whole of the world and that without any action from even a national institution the data in these formats will be accessible for another 10 years at least.

At this point we return to our goal; to keep material in good stewardship for the next generation of systems. If all our transferred data conforms to this pattern we will not need to intervene (in terms of transforming the data or translating it for our customers) as the material is already in a form that can be readily understood using ubiquitous internet or home computer facilities. We do not expect this situation to change during the life of the new DRI!

Our Challenges

The National Archives is faced by a number of significant challenges in terms of its work on digital preservation. While, even with the change from a 30 to 20 year rule (16), the first significant wave of digital records transferring from government departments is still some time away, other sources of digital records and other digital assets in need of preservation are already appearing. Some of these collections are in immediate danger of loss as the institutions that are creating them are themselves short lived.

New collections arriving

These collections (and indeed our view of the records we expect to see from government departments) are of a different shape to what was believed to be the situation when our earlier digital preservation systems were constructed. Our earlier system expected relatively small collections of records in a wide variety of “strange” formats. What we now see emerging are very large collections of records in a small number of common formats (in the case of digitised data vast collections in a single format). Our earlier system assumed the necessity for significant human oversight and intervention in the curation and processing of records, it is now abundantly clear that such an approach is both unnecessary and unsustainable; hence the creation of our new DRI.

We are seeing very large collections of digitised material (both as digital records, where the paper original is not available and as digital surrogates where the paper record exists but would be too costly or fragile to re-digitise). We have one of the world’s largest publically accessible Web Archive collections in the UK Government Web archive that contains over 1.5 Billion web pages that we need to secure and preserve for the long term. We now expect relatively large collections of “administrative” documents form government departments in common and familiar formats.

Institutions delivering records now

In terms of short lived institutions, all significant public enquiries and inquests now operate online, collating and recording their evidence, deliberations and conclusions digitally; the 2012 London Olympic Games was planned managed and delivered digitally and we are the archive that will preserve these records for the future. These types of institutions are beginning to deliver large collections of records to The National Archives. This can create issues which in the past the passage of time would have resolved. For example very up to date data often has very particular sensitivities that is no longer present in older data (one of the reasons for the 30 – soon to be 20 - year rule). Also, the passage of time often allows a more sober reflection on the historical significance of records being selected, which is not necessarily possible very soon after an apparently momentous event.

Wider context

The National Archives is at a “tipping point” between paper and digital. Our digital collections (in particular the UK Government Web Archive) by some reckonings now equal our paper holdings (~1.5 billion web pages compared to ~ 1 billion sheets of paper in our paper repositories). Unfortunately the there is a further 20 years worth of paper records still to come from government at the same volume as before so we cannot simply shift our attention to the digital. We have to operate more smartly, reduce complexity in our systems, do the minimum necessary and appropriate, and focus on the goal; in other words we have to apply the principle of parsimony.

Our Parsimonious Response

I will now go on to describe in broad terms the new DRI system we are building and refer back as I do so to the sections above to illustrate how parsimonious preservation is enabling us to respond to the challenges we face.

DRI – Overall structure

In designing the DRI system we have started from the point of view of the “processes” and “outcomes” that the system has to support and implement. We have focused on 5 aspects and the structure of our new system reflects these aspects. These aspects are:

- Transfer and Preparation
- Implementing the concept of “Safe Custody”
- Ingest and Accessioning
- Storage considerations
- Export

In other words we are focusing on the goal!

Transfer and Preparation

In my 2009 paper to this conference (1) I took the view (repeated above) that the failure to capture digital material is the biggest single risk to its preservation; our work at The National Archives continues to point to and highlight this risk. In our work this manifests itself in difficulties in the “Transfer Process” when supplier organisations attempt to provide us with material to preserve.

In the last year or two we have begun to apply the principle of Parsimony to this part of our work and in particular we are working to drastically simplify both the form and content of the metadata we require. For example our current plan (which will be finalised in guidance published on our website soon) for transfers from government departments focuses on only 6 metadata fields for each digital object transferred.

1. Title - A meaningful folder or file name.
2. Identifier - Not a system-generated ID number, but the filepath which supplies context for the record indicting its relationship to the activities of the organisation; this may be a URI or file-plan classification
3. Date - Note: this is not the date the record was copied to its current location; it should be the last date it was modified
4. Checksum - Generated using the SHA2 (256) algorithm. This guarantees the file has not changed during transfer.
5. Copyright – this will often be Crown Copyright

6. Closure status – This describes any sensitivity that may require the record to be closed for an extended period (e.g. the record contains personal data relating to a living individual)

We have also recognised that the nearer that supplying organisations can get to providing data and metadata “right first time” in its structure and format the faster and more efficient both ours and their activities will be.

With this in mind we encourage our suppliers to use our DROID tool (17) to examine the digital material they are considering for transfer. This enables them and us to determine as early as possible if there is material that is of high value that falls into the file format long tail and thus may require additional attention. We have developed a simple tool that enables suppliers to automatically generate 5 of the 6 fields we require directly from records in a shared folder file system and we are in the process of developing a tool that will enable suppliers to check for themselves that their submissions meet all of our metadata and format requirements.

In our systems we also aim to carry out comprehensive checks very early in our processes so that any problems can be addressed as quickly and as comprehensively as possible before our or our supplier’s time has been unduly wasted.

A quick reaction capability

The risk to loss of data from short lived institutions is another area where vigorous application of the principle of parsimony is already delivering benefits in enabling the safe capture and custody of significant records. We may not have time to configure our full processing capability to accession new types or structures of records between the conclusion of an institutions work and it being disbanded. Therefore, we have identified a set of minimal actions that permit us to confidently accept records, and take them into our safe custody, in the knowledge that we can accession the collection in slower time.

So what are these minimal actions? They are based on the observation that a parsimonious approach to digital preservation can be summed up in two lines

- Know what you have got
- Keep the bits safe

And that the first of these can be divided into a further two points

- Understand the file formats
- Catalogue your data

To achieve this we are deploying and developing a number of tools to process each batch of supplied data.

First we carry out two independent virus checks. Then to produce a rapid inventory of a batch of transferred data and confirm the absence of unwanted or unexpected file formats we use our own DROID tool.

To augment this and give is confidence in our ability to process the data in full we are developing a “metadata check” tool that will take a metadata template and establish that:

- the metadata supplied conforms to the template,
- all files supplied have an associated line of metadata,
- all metadata lines supplied refer to one and only one file supplied
- the hash value in the metadata line conforms to the hash value calculated for the file supplied

Finally we have developed a capability that encapsulates the batch of data, and stores three copies on to a managed tape archive. The three copies are saved to two distinct media types - one copy on tape designed for enterprise access and two copies on media designed for backup – one of the backup media copies is then stored securely off site as a disaster recovery measure.

Initially we expect these tools to be deployed and used manually; in the future we will develop a more automated environment to streamline this work.

Ingest for Accessioning

The main focus of this part of our system is the transformation of the metadata and record structures we receive to enable them to be listed in our overall catalogue and to render the material both safely preserved and discoverable.

This involves a finer grain processing of the data than we need to secure safe custody in the quick reaction capability described above. This processing creates a structure that enables the secure and safe storage of the material together with properties needed to present the material to customers (in particular an identification of the file format so that the customer knows which software to use to view the data).

The structure also allows for the data to be re-exported on demand in the form in which it was originally presented to the system. This fulfils another parsimonious principle that the system makes no irrevocable changes to the data and that as a "steward", the data can be passed on to a subsequent system in good order

Storage

It is well known in the storage engineering community that media failure is not random and that it largely arises from errors in manufacturing, in a very clear paper (5) Rosenthal argues that

"Practical digital preservation systems must therefore:

- *Maintain more than one copy by replicating their data on multiple, ideally different, storage systems.*
- *Audit or (scrub) the replicas to detect damage, and repair it by overwriting the known-bad copy with data from another."*

We use a managed tape archive with exactly these properties: diverse media from different manufacturers and media monitoring to detect non catastrophic read errors which lets us correct or replace the faulty media detected.

This tape archive has the added benefit of significantly reducing our carbon footprint and, in the knowledge that the master copies rarely require access, reflects an added benefit of parsimony in terms of cost compared to the unnecessarily functional disk based alternative.

In a further application of parsimony, we do not preserve presentation copies since, as we know the master material is not obsolete, we could re-generate the presentation copies if they were to become lost. Our presentation systems are of course managed using good IT practices including robust data backup so the likelihood of needing to regenerate even these copies is very low.

Something to Note

In all of the above discussion readers familiar with digital preservation literature will perhaps be surprised not to see any mention or discussion of “Migration” vs. “Emulation” or indeed of “Significant Properties”. This is perhaps one of the greatest benefits we have derived from adopting our parsimonious approach – no such capability is needed! We do not expect that any data we have or will receive in the foreseeable future (5 to 10 years) will require either action during the life of the system we are building. A truly thrifty outcome!

Conclusion

I hope that I have demonstrated that the principle of Parsimonious Preservation originally developed in 2009 as an approach for small or medium sized institutions to permit them to begin work on digital preservation is also practical for large scale institutions. My aim in clarifying the core of digital preservation to the two steps “know what you have got” and “keep the bits safe” and in showing you how we at The National Archives are approaching our challenges, is to convince you that an appropriate response to digital preservation is indeed “Don’t Panic!”; or perhaps “Keep Calm and Carry On”.

Bibliography

1. **Gollins, Tim.** Parsimonious Preservation: Preventing Pointless Processes. *The National Archives.* [Online] December 2009. [Cited: 12 November 2012.] http://www.nationalarchives.gov.uk/documents/information-management/parsimonious-preservation.pdf.

2. **The National Archives.** Digital preservation FAQs, Technology. *the National Archives.* [Online] 18 October 2012. [Cited: 12 November 2012.] http://www.nationalarchives.gov.uk/information-management/projects-and-work/technology.htm.

3. **Merriam-Webster Online Dictionary.** parsimony. *Merriam-Webster Online Dictionary.* [Online] [Cited: 12 November 2012.] http://mw1.meriam-webster.com/dictionary/parsimony.

4. **The National Archives.** The National Archives Preservation Policy. *The National Archives.* [Online] 12 June 2009. [Cited: 12 November 2012.] http://www.nationalarchives.gov.uk/documents/tna-corporate-preservation-policy-2009-website-version.pdf.

5. **Rosenthal, David.** Keeping bits safe: how hard can it be? *Commun. ACM.* 2010, Vol. 53, 11, pp. 47--55. http://doi.acm.org/10.1145/1839676.1839692.

6. —. Presentations from the PASIG Summer Meetings in Malta. *Stanford Universtity Libraries and Academic Information Resources.* [Online] 24 - 26 June 2009. [Cited: 13 November 2012.] http://lib.stanford.edu/files/rosenthal_pasig_lockss.pdf.

7. **Mitchell, Joni.** Big Yellow Taxi (lyrics). *Roots of Bob Dylan.* [Online] 1970. [Cited: 12 November 2012.] http://www.bobdylanroots.com/bigyellow.html.

8. **Answers.com.** First catch your hare. *Answers.com.* [Online] 2012. [Cited: 12 November 2012.] http://www.answers.com/topic/first-catch-your-hare.

9. **The National Archives.** Records collection policy. *Plans, policies, performance and*

projects > Our policies > Draft policies . [Online] 16 July 2012. [Cited: 12 November 2012.] http://www.nationalarchives.gov.uk/documents/draft-records-collection-policy-2012.pdf.

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

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

12. **Rosenthal, David.** Format Obsolescence In The Wild? *DSHR's Blog.* [Online] 8 November 2012. [Cited: 13 November 2012.] http://blog.dshr.org/2012/11/format-obsolescence-in-wild.html.

13. —. Formats through time. *DSHR's Blog.* [Online] 9 October 2012. [Cited: 13 November 2012.] http://blog.dshr.org/2012/10/formats-through-time.html.

14. —. Cleaning up the "Formats through tIme" mess. *DSHR's Blog.* [Online] 13 October 2012. [Cited: 13 November 2012.] http://blog.dshr.org/2012/10/cleaning-up-formats-through-time-mess.html.

15. **Jackson, Andrew.** Formats over Time: Exploring UK Web History. *Cornell University Library.* [Online] 5 October 2012 . [Cited: 13 November 2012.] presented at iPres 2012 in Toronto. http://arxiv.org/pdf/1210.1714v1.pdf.

16. **The National Archives.** 20 Year Rule. *The National Archives, About Us, Our Projects.* [Online] 11 July 2012. [Cited: 16 November 2012.] http://www.nationalarchives.gov.uk/about/20-year-rule.htm.

17. —. DROID download page. *DROID (Digital Record and Object Identification), The National Archives, Digital Preservation.* [Online] March 2013. [Cited: 7 March 2013.] http://www.nationalarchives.gov.uk/information-management/projects-and-work/droid.htm.