DJ32004 RESEARCH & CREATIVE PRACTICE PART 1
LITERATURE REVIEW

DIGITAL PRESERVATION

Victor Loux
April 2014

Digital Interaction Design
University of Dundee

Table of contents

2 Table of contents

3 Introduction

4 The need for digital preservation

6 Threats to digital continuity

7 Who must preserve, and what we should keep

7 How to preserve

8 Methods for software preservation

9 Choosing a method

9 Personal archiving

12 Conclusion

13 Bibliography

17 Project Proposal Form

Introduction

As we shift into the Information Age and a knowledge-based economy, a number of concerns are being raised over the long-term preservation of digital data and artifacts.

This literature review aims to get an overview of the current problems, techniques and practices in preserving data on digital media, including the definition and understanding of the new paradigms created by digital preservation, the importance of being able to store things on the long term, the processes used to curate and select content to be preserved, and the various methods used to do so. It also encompasses personal archiving which will be the focus of my own research.

The need for digital preservation

With an exponential growth in data being produced digitally (SINTEF, 2013), comes a significant growth in the volume of data worth archiving over the long term (Harvey, 2012, p. 33). However, the preservation of digital materials cause much more issues than traditional preservation paradigms; where we can consider that for paper and physical artefacts, "benign neglect may be the best treatment" (Harvey, 2012, p. 10; derived from Bastian, Cloonan and Harvey, 1993; 2011), the rapid changes in technology and the fragility of digital media make this rule dangerous for the survival of data.

First, we should define what are archives, and what they are useful for. As Cook eloquently puts it:

> "Archives [...] are a source of memory about the past, about history, heritage, and culture, about personal roots and family connections, about who we are as human beings and about glimpses into our common humanity through recorded information in all media, much more than they are about narrow accountabilities or administrative continuity." (Cook, 2000)

It is certainly difficult to understand which data was already lost, and what is at stake now. Harvey emitted some criticism of pessimistic views for archiving:

> "It was suggested in 2002 that the last 25 years have been a 'scenario of data loss and poor records that has dogged our progress' and that, if this is not reversed, 'the human record of the early 21st century may be unreadable' (Deegan and Tanner, 2002). Rhetoric of this kind is common in the literature, but is regrettably poorly supported with specifics and evidence. Alarmist descriptions abound: there will be 'a digital black hole... truly a digital dark age from which information may never reappear' (Deegan and Taylor, 2002) if we do not address the problem, and we will become an amnesiac society. [...] It is only possible to conclude, as the authors of the Digital Preservation Coalition's handbook (2008, p. 32) did for the UK, that the evidence of data loss is 'as yet only largely anecdotal ... [but] it is certain that many potentially valuable digital materials have already been lost'." (Harvey, 2012, pp. 33–34)

While it is difficult to argue that preserving digital objects is necessary, there is certainly a need to make a selection of what deserves to be protected, and to what level. It is clear that the long term preservation of legal and governmental records, scientific publications (Harvey, 2012, p. 26) and business records (Ross and Gow, 1999, p. iii) on digital form is crucial; but the archival of the wider Internet, cultural artefacts and our personal heritage is subject to discussion, as fewer items could be considered valuable.

UNESCO note that those who value a more comprehensive collection argue that "any information may turn out to have long-term value, and that the costs of detailed selection are greater than the costs of collecting and storing everything" (UNESCO, 2003, p. 73). Additionally, digital storage media degrade so rapidly that the content might disappear before a decision about its value has been made (UNESCO, 2003, p. 71; Harvey, 2012, p. 57). Proponents of a more selective collection argue that we could obtain a higher content quality, settle the preservation rights with content producers, and that we do not have the physical resources to store everything anyway (UNESCO, 2003, pp. 70-73; Harvey, 2012, pp. 26, 58).

Both UNESCO (2003) and Cook (2000) recognise that a middle ground must be found, and that both approaches are valid depending on the goal of the preservation intent. To give an example in archiving the Web, the Internet Archive (Kahle, 1996) has chosen to store copies of as many websites as possible, but due to technical limitations discards all images, stylesheets and multimedia objects; evidently this is not appropriate for valuable websites. The British Library's Web Archive[1] curates collections of websites by topic and by events (Queen Jubilee, London 2005 terrorist attacks, General Elections, Olympic and Paralympic Games, ...) to reflect the state and opinion of the Web at these given points in history; because fewer sites are archived and their content is deemed important, they have archived full quality versions of these sites. See Masanès (2006, p. 83) for a methodology for this type of selection. Cook reminds us however that archivists have a duty of impartiality and objectivity during the selection and must ensure all points of view are being kept, to ensure our shared memory does not "becomes counterfeit, or at least transformed into forgery, manipulation, or imagination" (Cook, 2000).

Hiroyuki Kawano also proposes to use reputation models to let crawling robots consider the

1 http://www.webarchive.org.uk/ukwa/ [Accessed: 11 Apr 2014]

"importance, fairness, trustiness, uniqueness and valuation" of a Web page (Kawano, 2008) based on numerous parameters, to decide whether a page is worth archiving. The automated nature of this selection can, by nature, be inaccurate and inadvertently subjective, therefore a manual selection is still sometimes required.

Threats to digital continuity

UNESCO (2003, pp. 30–31) has compiled a list of threats to what they call *digital continuity*. The two principal ones are the short lifetime of digital materials carriers (i.e. storage media), and the obsolescence of the means to access these digital materials (i.e. the software to read files and the hardware to read discontinued storage media). Research by the Digital Preservation Coalition (2008, p. 154) show that the reliability of storage material varies highly depending on the type of material and the storage conditions, going from a few months to decades or even centuries.

But the media are only part of the equation; because all storage media eventually become obsolete, we will not necessarily be able to read them again even if they are not deteriorated. Weathley (2003, cited by Harvey, 2012, p. 37) denounced that myth of long-lived media: "[an IT vendor] offered us a special polymer that they guaranteed would preserve a CD-ROM for 100 years. They were unable to answer how we would preserve a CD-ROM player for that length of time." And as Pearson adds, not only is the CD-ROM drive needed, but also the type of cable to plug that drive to the computer, a computer motherboard that has a connector for this cable, an operating system with drivers for these types of connectors, and a program able to open the file itself (Pearson, 2009, slide 11). All of these elements become rapidly obsolete themselves, which mean that media may not be easily accessible on current technology after just one or two decades, even if the storage media last longer than that. Therefore it is needed to periodically migrate (a process also called "refreshing") the bitstreams from a media type to a more contemporary one in order to ensure that this bitstream will remain easily accessible (Harvey, 2012). As the bitstream is preserved, only the question of preserving software to read it is left.

Who must preserve, and what we should keep

Digital media also changes the original preservation paradigm for deciding who takes the responsibility for archiving content, and how. The processes of creation, fabrication and publication used to be distinct in the "analogue" era. But as Nurnberg (1995, p. 21, cited in Harvey, 2012, p. 9) said, "technology tends to erase distinctions between the separate processes of creation, reproduction and distribution that characterise the classic industrial model of print commodities". Self-publication (particularly on websites) is much more common, and for many publishers the task of preservation is new to them (Harvey, 2012, p. 32); they might not have the awareness, knowledge or resources to do so themselves it is therefore essential that they collaborate with preservation organisations, such as libraries and archives (Ayre and Muir, 2004). See Harvey (2012, chapter 9), Ayre and Muir (2004) and UNESCO (2003, chapter 11) for more resources and case studies.

Ayre and Muir (2004) suggest that the complexity and resources needed for digital preservation may mean that individual libraries may not have the possibility to do the preservation themselves like they used to, but instead rely on national libraries and centralised preservation systems to reduce overhead. This system is used by a number of libraries; even here at the University of Dundee, most journals and a selection of books are available through centralised systems (Dawsonera, Ebrary) or publishers directly (ACM, Elsevier Science, Springer).

How to preserve

There are multiple methods for preservation. Each has advantages and issues, although one parameter is particularly influential: **authenticity**. This refers to the faithfulness of the rendition of the preserved object in the future, compared to when it was created; while the content should not be altered, its presentation may change (Cook, 2000). What is considered appropriately authentic is subject to debate (del Pozo *et al.*, 2012, p. 7; Cook, 2000) and may vary depending on the preservation intent; del Pozo *et al.* (2012, p. 8) give, as example, the normalisation of a spreadsheet to formats that may accurately retain either the formatting, or the

formulae used in each cell.

Retaining integrity, that is ensuring the object is "what it purports to be" and is "complete and has not been altered or corrupted" (Ross, 2002, p. 7, cited in Harvey, 2012, p. 54), can also be crucially important, particularly for proving that a business or legal document has not been tampered with (Harvey, 2012, p. 54). But as Lynch (2000), UNESCO (2003, p. 22) and Harvey (2012, p. 54) argue, integrity is mostly a matter of trust and can only be ensured with a thorough documentation of the archiving process.

Methods for software preservation

[NB: due to the length limit of this literature review, I only give a simplified overview of the main archiving techniques. See Granger (2000), Thibodeau (2002) and Harvey (2012) for a comprehensive list of methods, including configurable chips, persistent archives, object interchange format, etc.]

- **Hard copies**: print documents to archive them. Rarely considered a viable option for large data sets and interactive documents (Granger, 2000; Rothenberg, 1998).
- **Computer Museums**: Swade (1993) proposed to preserve hardware and software centralised places. Considered short-term and prohibitively expensive by Granger (2000) but can be useful to recover old media or testing emulators (Rothenberg, 1998).
- **Migration**: the most popular option, consisting of periodically saving files to a current file type (newer version of a format, different format, or a standardised format) (Granger, 2000; Thibodeau, 2002) to avoid technological obsolescence. It is labour intensive and risky, as it compromises authenticity (Rothenberg, 1998).
- **Universal Virtual Computer (UVC)**: proposed by Lorie (2000, in Thibodeau, 2002, p. 22). It consists of implementing a format viewer in a portable programming language like Java[2] which could then theoretically run on any future platform supporting that programming language. It is however slow and functionally limited (Lorie 2000), and runs the risk of the portable language itself becoming obsolete.

2 Most programming languages include instructions that are specific to a given platform, therefore the same file will need to be adapted to run on another platform. "Portable" languages like Java are called high-level and are more abstract. They are run inside a virtual machine (itself low-level, different for every platform) which worries about transforming Java code into instructions that are specific to that platform; as such, the same file of code can be ran on any platform.

- **Emulation** is the most conservative model; it simulates a legacy architecture on current platforms, allowing to run the original software used to create a file, therefore not modifying them. It "keeps the look and feel as well as the interactivity" (Granger, 2000) therefore is highly relevant for video games (Guttenbrunner *et al.*, 2010) and interactive systems, and it guarantees authenticity. It is however highly complex to implement (Granger, 2000; byuu, 2011) and need renewal, since emulators themselves become obsolete (Thibodeau, 2002), as exemplified by emulators provided by Apple for their architecture switches (Mesa, A. F., 1997; AppleInsider, 2011) that were later discontinued, effectively removing the ability to run pre-2005 applications on their current models.

Choosing a method

Thibodeau (2002, pp. 15–16) suggests four criteria to choose a preservation method. It must be:

- technically feasible
- sustainable and resistant to technological obsolescence
- practical (the preserving organisation must have the resources to do it)
- appropriate to the type of material being preserved, and the objectives of the preservation (see del Pozo *et al.*, 2012, p. 7 and Granger, 2000 for examples).

Personal archiving

It is interesting to ask ourselves if the paradigms for digital preservation used by librarians and archivists are still valid for preservation of our personal data. Harvey (2012) and Lukesh (1999, cited in Harvey, 2012, p. 32) note that our personal correspondence and sentimental artefacts stored digitally might be lost if we do not take steps to preserve them, and wonder how we will be able to understand our modern life without these exchanges.

While this is a correct idea, I think that few people kept their letters in the sole purpose of leaving researchers decades from now understand our epoch's social habits and ways of speaking. There has to be other motivations for preserving our conversations and our objects.

Richard Banks, a researcher at Microsoft Cambridge, suggests that we first preserve artefacts for ourselves. The majority of the time, we simply keep what we consider significant to remind ourselves of particular people, events and places, and to share them with people we care about; we discard what is not meaningful to us by pragmatism (lack of physical space, lack of importance, redundancy) (Banks, 2011, p. 6). This selection process contributes to keeping what we want to remember, and what is the legacy we will want to leave to others after our deaths.

However, the digital age changes this situation. First, the space to store our objects on computers and online is becoming virtually unlimited, and so we tend to accumulate data instead of selecting (Banks, 2011); second, the abundance of online services makes it harder to keep track of everything we create. Fortunately, digital objects have fundamental properties that physical artefacts don't have, which actually provide numerous advantages in terms of preservation: the ease of organising a collection of data through indexing, cataloguing and adding metadata (Kirk and Sellen, 2010, p. 35), and the ability to duplicate and share that object accurately and virtually cost-free (Banks, 2011).

The cataloguing advantage allows us to process it using computer science. Hangal *et al.* (2011) have developed MUSE (Memories USing Email), a system finding patterns in email archives to help us make sense of this large data set. As the authors note in the introduction, "email has become a de facto medium of record; many people consciously deposit important information into email, knowing they can look it up later", and so our inboxes contain highly valuable information.

Early users reported that they have been using it to: make a summary of their work progress; extract and organise certain type of data (for instance, the personal out of the professional, or the important out of the more mundane); finding life milestones inside family emails; picking up work that has been left unfinished and forgotten; or renewing with old relationships. These uses went beyond the authors' original expectations of simply reminiscing:

> "The stories above include an example of each one of the '5R's' described by Sellen and Whittaker [2010]: recollection, reminiscing, retrieving, reflecting, and remembering intentions. Further, it suggests that browsing and remembering the past can affect the future." (Hangal *et al.*, 2011)

The copying advantage, while useful for sharing, has issues for backing up. Banks (2011) notes the process is still troublesome, and still *feels* insecure because of cheap looking media, and nowadays, the abstraction with cloud storage (about Flickr: "I have no idea what kind of hardware my files are now stored on or even where they are geographically. I just expect to have access to them as long as I pay my bills" Banks, 2011, p. 27).

Kirk and Sellen (2010) found however that ultimately, stories and narrative are more important than the preservation of the artefact itself. They propose that "technology might play a role in capturing and associating stories or narratives with different physical objects", imagining we could also digitally augment physical objects or places. Banks (2011) gives as example the Weather Camera (Wilkens, 2011), which records wind force and temperature along with commentary; a system of physicals backup of sentimental items by Serrano (2009) using 3D scanning and 3D printing; and tools by Microsoft Research to recreate an environment in 3D just using photos. A similar technique is used in bigger preservation projects to make high quality 3D models of culturally important places and objects in China (Zhou *et al.*, 2012).

Conclusion

This literature review, although shallow considering the breadth of the topic, allowed me to gain an understanding of the current practices and debates in the digital archiving community, and to gather relevant sources for any future research in each sub-topic. I found the theme of personal archiving captivating but not explored as intensely as large scale preservation for our common knowledge and data; I would like to do deeper research in this particular area of digital preservation, notably on our use of social media and Internet services on the long term.

Bibliography

AppleInsider. (2011) '*Inside Mac OS X 10.7 Lion: Missing Front Row, Rosetta and Java runtime*'. *AppleInsider*. [web page] Accessible at http://appleinsider.com/articles/11/02/26/mac_os_x_lion_drops_front_row_java_runtime_rosetta.html [Accessed: 19th Apr 2014]

Ayre, C. and Muir, A. (2004) 'The right to preserve'. *D-Lib Magazine*, vol. 10 (No. 3). March. Available at http://www.dlib.org/dlib/march04/ayre/03ayre.html [Accessed: 11th Apr 2014]

Banks, R. (2011) *The Future of Looking Back*. Sebastopol: Microsoft Press.

Bearman, D. (1999) 'Reality and Chimeras in the Preservation of Electronic Records'. *D-Lib Magazine*, vol. 5 (No. 4). April. Available at http://www.dlib.org/dlib/april99/bearman/04bearman.html [Accessed: 7th Apr 2014]

Boyd, D. (2014) *It's Complicated: The Social Lives of Networked Teens*. New Haven and London: Yale University Press.

byuu.[1] (2011) 'Accuracy takes power: one man's 3 GHz quest to build a perfect SNES emulator'. *Ars Technica*. [web page] Available at: http://arstechnica.com/gaming/2011/08/accuracy-takes-power-one-mans-3ghz-quest-to-build-a-perfect-snes-emulator/ [Accessed: 8th Apr 2014]

Cook, T. (2000) 'Beyond the screen: the records continuum and archival cultural heritage.' Presented at the *Australian Society of Archivists Conference*, Melbourne, 18 August. Available at http://www.mybestdocs.com/cook-t-beyondthescreen-000818.htm [Accessed: 9th Apr 2014]

Granger, S. (2000) 'Emulation as a Digital Preservation Strategy'. *D-Lib Magazine*, vol. 6 (No. 10). October. Available at http://www.dlib.org/dlib/october00/granger/10granger.html [Accessed: 7th Apr 2014]

Guttenbrunner, M., Becker, C. and Rauber, A. (2010) 'Keeping the game alive: Evaluating strategies for the preservation of console video games'. *International Journal of Digital Curation*, vol. 5 (Iss. 1), pp. 64-90.

1 The article is only credited to the developer's pseudonym.

Bibliography (cont.)

Hangal, S., Lam, M. S. and Heer, J. (2011) 'MUSE: reviving memories using email archives', in *Proceedings of the 24th annual ACM symposium on User interface software and technology (UIST '11)*. New York, NY: ACM, pp. 75-84. doi:10.1145/2047196.2047206

Harvey, D. R. (2012) *Preserving digital materials*. Berlin: De Gruyter Saur.

Kahle, B. (1996) 'Archiving the Internet'. *Scientific American*, Iss. March 1997. Available at http://web.archive.org/web/19971211123138/www.archive.org/sciam_article.html [Accessed: 27th Feb 2014]

Kawano, H. (2008) 'Strategy of Digital Contents Archive Based on Reputation Model', in *ICSENG '08. Proceedings of 19th International Conference on Systems Engineering*, 19-21 Aug, pp. 288-293. doi: 10.1109/ICSEng.2008.75

Kirk, D. S. and Sellen, A. (2010) 'On human remains: Values and practice in the home archiving of cherished objects'. *ACM Transactions on Human-Computer Interaction,* vol. 17 (Iss. 3), article 10. doi:0.1145/1806923.1806924

Library of Congress. (2007) *Digital Preservation Program Makes Awards to Preserve American Creative Works*. [press release] Available at http://www.loc.gov/today/pr/2007/07-156.html [Accessed: 14th Apr 2014]

Lubar, S. (1999) 'Information Culture and the Archival Record', in *American Archivist*, vol. 62 (Iss. 1), pp. 10-22.

Lynch, C. (2000) "Authenticity and Integrity in the Digital Environment: An Exploratory Analysis of the Central Role of Trust', in *Authenticity in a Digital Environment*. Washington, DC: Council on Library and Information Resources. Available at http://www.clir.org/pubs/abstract/pub92abst.html [Accessed: 16th Apr 2014]

Masanès, J. (2006) *Web archiving*. Berlin: Springer.

Mesa, A. F. (1997) *'The PowerPC Triumph'. The Apple Museum*. [web page] Available at: http://applemuseum.bott.org/sections/ppc.html [Accessed: 19th Apr 2014]

Bibliography (cont.)

Paloque-Bergès, C. (2011) *Entre trivialité et culture : une histoire de l'Internet vernaculaire. Émergence et médiations d'un folklore de réseau* [*Between Triviality and Culture: a History of the Vernacular Internet. Emergence and mediations in network folklore*]. PhD thesis, self-published. Available at http://camillepaloqueberges.wordpress.com/phd-thesis/ [Accessed: 6th Apr 2014] [2]

Paloque-Bergès, C. (2013) 'Un patrimoine composite : le public Internet face à l'archivage de sa matière culturelle' [A composite heritage: the Internet public confronted to the archiving of its cultural matter], in I. Dragan, P. Stefanescu, N. Pelissier, J.-F. Tétu and L. Idjeroui-Ravez (eds.) *Traces, mémoire, communication*. Bucharest: Editura Universității din Bucureşti. [2]

Pearson, D. (2009) *Preserve or Preserve Not, There is No Try: some dilemmas relating to Personal Digital Archiving*. [PowerPoint slides]. National Library of Australia Staff Papers. Available at http://www.nla.gov.au/openpublish/index.php/nlasp/article/view/1388 [Accessed: 9th Apr 2014]

del Pozo, N., Long, A. S. and Pearson, D. (2012) 'Land of the Lost: a discussion of what can be preserved through digital preservation'. *Library Hi Tech*, vol. 28 (Iss. 2), pp. 290-300. Available from: doi: 10.1108/07378831011047686. [Accessed: 6th Apr 2014]

Ross, S. and Gow, A. (1999) *Digital archaeology: rescuing neglected and damaged data resources: a JISC/NPO study within Electronic Libraries (eLib) Programme on the Preservation of Electronic Materials*. London: Library Information Technology Centre. Available at http://www.ukoln.ac.uk/services/elib/papers/supporting/pdf/p2.pdf [Accessed: 17th Apr 2014]

Rothenberg, J. (1998) *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation*. [report] Washington, DC: Council on Library and Information Resources.

2 In French.

Bibliography (cont.)

Serrano, H. (2009) *Back Up Objects, Héctor Serrano*. [web page] Available at: http://www.hectorserrano.com/index.php?id=41&m=lab&grupo=backup [Accessed: 17th Apr 2014].

SINTEF. (2013) *'Big Data, for better or worse: 90% of world's data generated over last two years.' ScienceDaily.* [web page] Available at: http://www.sciencedaily.com/releases/2013/05/130522085217.htm [Accessed: 16th Apr 2014]

Swade, D. (1993) 'The problems of software conservation", in *Computer Resurrection*, vol. 7. Available at: http://www.cs.man.ac.uk/CCS/res/res07.htm#f [Accessed: 12th Apr 2014]

Task Force on Archiving of Digital Information. (1996) *Preserving digital information, Report of the Task Force on Archiving of Digital Information*. Washington, DC: Council on Library and Information Resources.

Thibodeau, K. (2002) 'Overview of technological approaches to digital preservation and challenges in coming years', in *The state of digital preservation: an international perspective*. Washington, DC: Council on Library and Information Resources, pp. 4-31.

UNESCO. (2003) *Guidelines for the preservation of digital heritage*, prepared by the National Library of Australia. Paris: UNESCO. Available at http://unesdoc.unesco.org/images/0013/001300/130071e.pdf [Accessed: 9 Apr 2014]

Wilkens, K. (2011) *Sensor Poetics,* Kjenwilkens.com. [web page] Available at http://www.kjenwilkens.com/projects/sensor-poetics [Accessed: 18th Apr 2014]

Zimmer, M. (2010) '"But the data is already public": on the ethics of research in Facebook', in *Ethics and information technology*, vol. 12 (Iss. 4), pp. 313-325. doi: 10.1007/s10676-010-9227-5

Zhou, M., Geng, G. and Wu, Z. (2012) *Digital preservation technology for cultural heritage*. Berlin: Springer.

Project Proposal Form

Name	Victor Loux
Programme/Course	Digital Interaction Design
Project/working title	Digital preservation for the masses
Project Aims (i.e. what you want to learn about through further investigation. E.g. I want toa learn more about the creative process of animators)	I would like to look at further research papers on personal preservation and focus on ways to apply "large-scale" paradigms to our personal heritage while trying to be respectful of people's privacy and unintrusive.
Objectives (i.e. the particular things you propose to do in order to carry out your investigation. E.g. Make contact with professional animators and arrange to meet them. Possibly Interview them or observe their creative activities to gather data)	I first intend to do further reading in this area, as until now I have been exploring digital preservation as a whole with no particular focus, to get a broader understanding of the issues and solutions. I would like to interview several people of different age ranges to understand their relationship to technology and their personal digital heritage, notably concerning photographs and conversations.
Methods (i.e. the different ways in which you will achieve these objectives. E.g. Open ended interview techniques and grounded theory analysis or video observation of working animators)	Using open ended interview techniques, further literature review and possibly research through practice as I intend to explore this theme for my final year project.