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 legal and ethical issues arising from it, and how the paradigms and methods change when applied to different types of archiving (e.g. for researchers, for businesses, and for personal businesses).

Why digital preservation

[Stats of numbers on digital vs physical]

Why digital degrades faster

UNESCO's Threats to digital continuity

UNESCO (2003, pp. 30–31) has identified a list of threats to what they call *digital continuity*. These are: + "The carriers used to store these digital materials are usually unstable and deteriorate within a few years or decades at most". Where, in the traditional preservation paradigms for paper-based artifacts, the main rule is that "benign neglect may be the best treatment" (derived from Cloonan (1993, p.596), Harvey (1993, pp.14,140), and Bastian, Cloonan and Harvey (2011, pp.612–613)), this cannot be applied to digital storage media because their lifetime is much shorter. This varies highly with the type of medium used and with the storage conditions; for example, samples by the Digital Preservation Coalition (2008, p. 154) show that D3 cartridge tapes deteriorates after 50 years in optimal conditions (25% RH at 10°) but only one year in high humidity and temperatures (50%RH at 28°). Similar results happen for the more commercially available optical media (down to three months and up to 200 years), which are more sensitive to moisture (Harvey, 2012, p.47).

- "Use of digital materials depends on means of access that work in

particular ways: often complex combinations of tools including hardware and software, which typically become obsolete within a few years and are replaced with new tools that work differently" (UNESCO, 2003). Even if storage media could survive decades or even centuries, technological change means we will not necessarily have the means to access the data later. In other words, "to retain appropriate access to a digital object, we must maintain a mechanism to derive meaning from that object" (del Pozo et al., 2012, p6). for optical media, this includes not only the drives to read the disc, but also a cable to connect the drive to the computer's motherboard, a motherboard that has a connector for this type of cable, and an operating system that has drivers for this connectivity (Pearson 2009, slide 11). Once we have access to the bitstream from the computer we need to be able to represent it, which means we also need to keep the software in a version that can open the preserved file. There is a myth, kept by both customers and vendors, that media is long-lived. "[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." (Weathley, 2003, cited by Harvey, 2012, p. 37)

- Poor preservation documentation and metadata means it could be hard for potential users to find the data (UNESCO 2003)

-
  - poor preservation ("Critical aspects of functionality, such as formatting of documents or the rules by which databases operate, may not be recognised and may be discarded or damaged in preservation processing.", UNESCO 2003) + "So much contextual information may be lost that the materials themselves are un-intelligible or not trusted even when they can be accessed" -> poor integrity

- "Materials may be lost in the event of disasters such as fire, flood, equipment failure, or virus or direct attack that disables stored data and operating systems"

- "Access barriers such as password protection, encryption, security devices, or hard-coded access paths may prevent ongoing access beyond the very limited circumstances for which they were designed". [This argument is particularly true for web archiving, where crawlers of national libraries retrieve all sort of files to preserve them but do not

necessarily have the permission or know the rights owner][back that up maybe yo]

Importance of data

Library material: for society

> "if we librarians do not rise to the occasion, successive generations will know less and have access to less for the first time in human history. This is not a challenge from which we can shrink or a mission in which we can fail" (Gorman, 1997, cited in Harvey, 2012, p. 27)

For businesses (records)

Personal archiving / domestic archiving

Preserving interactive material inc. video games

In addition to the preservation of "static" documents and images, there is a significant desire to preserve video games, which are getting recognised as a form of art similarly to movies and music. The Library of Congress (2007) gave awards for the preservation of video games, there is a growing interest in being able to play "retro" games, but the consoles are obsolete. A large community of hobbyists reverse engineered consoles and produced emulators of these platforms for PC and other platforms different from their intended one. Because no documentation is available, this is a Herculean task, byuu (2011), the developer of an emulator for Super Nintendo games, recalls this for reengineering certain cartridges.

> [Low-level emulation] is also a very expensive operation, monetarily speaking: to obtain the DSP program code requires melting the integrated circuit with nitric acid, scanning in the surface of a chip with an electron microscope, and then either staining and manually reading out or physically altering and monitoring the traces to extract the program and data ROMs. This kind of work can cost up to millions of dollars to have done professionally, depending upon the chip's complexity, due to the extremely specialized knowledge and equipment involved. (byuu, 2011)

Obviously this is not necessarily a route many institutions would like to take for preserving digital objects. Because this also causes intellectual

property issues, Nintendo and Sony have launched their own preservation programs, allowing players to play "classic" games on the newer generations of consoles; this is a far more accurate emulation (as they do have the internal documentation to create these emulators) and respects their monetary interests (players have to buy the games, whereas in hobbyist emulation, you would have to download the game file).

Useful for researchers in the future

Beyond the governmental and cultural needs for preservation, there is also great need for preserving scientific data. "Data are the foundation on which scientific, engineering, and medical knowledge is built" (Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age, 2009, p. ix, cited in Harvey, 2012, p. 26); therefore it is crucial in a digital era, with enormous data sets and research papers becoming widely public, to ensure this doesn't disappear.

Understanding our history

There is a need to define what is the purpose of archives. As a tool to learn from the past, or to keep records, but as Cook (2000) 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)

Preservation should, therefore, be considered very important. There is some skepticism to this.

> “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’s difficult to argue that preserving digital objects is necessary, there is certainly a need to define what deserves to be protected, and to what level.

Understanding the beginnings of the Internet, as we have shifted in an information society

Preserving an accurate + [integrity] image of our society at a given point

Twitter & Usenet

Camille Paloque-Bergès (2011, 2013) proposes that we can already find historical and sociological evidence of a given period in Internet time when browsing archives of Usenet [footnote: An Internet discussion system established in the 1980s and that is widely regarded as the predecessor of the Web and online fora]. She analysed how the way of speaking has evolved [footnote: See also http://wwwtxt.org (or @wwwtxt on Twitter), a project collecting short phrases of the Usenet to demonstrate this] and how an self-organised community had built itself, noting that we can “feel [the community’s] informational generativity” and understand that one of the biggest qualities of the early users of the Internet was their “recursivity as a public” (Paloque-Bergès, 2013), that is, how they welcomed newcomers — notably through the use of FAQs [footnote: Frequently Asked Questions], which weren’t a common concept back then. There are also large elements of nostalgia coming back (Paloque-Bergès, 2011).

It is also interesting to come back on the archival of Usenet. The messages were not systematically archived, and obviously not available on the Web as

it did not exist and is separate. Today's collections have mainly been recreated from donations of people who did archived Usenet newsgroup, originally to DejaVu News in 1995, and Google Groups in 2001. A number of privacy concerns were raised with the publication of these searchable archives; messages who were, 15 to 20 years earlier, thought to be confidential are now considered as digital heritage. There was also a large backlash against the archiving made by for-profit companies, and the credit taken later by Google for its donors, with other critiques like Brad Templeton's (an Internet veteran and important figure, cited in Paloque-Berges, 2013) remark that the archive was lacking in integrity. However, the searchability of these archives did bring welcomed technical possibilities for an Internet historiography, including a timeline where "the announcement of the Chernobyl disaster on Usenet stands alongside the first message of Linus Torvalds calling for help on what would become the Linux operating system" (Paloque-Bergès, 2013).

Similar concerns appeared at the announcement of the archival of Twitter by the Library of Congress in 2010. The two are not really comparable, mostly because the archive is not public and is made by a not-for-profit body. On the mailing list of the Association of Internet Researchers, the idea was met with interest. "The public twitter stream is of historical cultural significance and is an amazing repository of mundane moments in the daily lives of many people and records of what they thought important" (Baym, 2010) [on http://listserv.aoir.org/pipermail/air-l-aoir.org/2010-April/021125.html]". However, the privacy concerns (notably with public and private feeds) were discussed, and as Michael Zimmer remarked:

> "This is the classic "but the information is already public" argument that, while technically true, presumes a false dichotomy that information is either strictly public or private, ignoring any contextual norms that might have guided the initial release of information or how a person expects that information to flow." (Zimmer, M. [on http://listserv.aoir.org/pipermail/air-l-aoir.org/2010-April/021136.html]; see also his paper (2010) on Facebook privacy, going further in that argument)

Paloque-Bergès (2010) theorises this argument, suggesting that "information is not an object, but a statement; it goes through socio-technical processes; it doesn't have an objective nature". A message in a tweet can have a very

different context when taken apart of the moment it was posted. It also emphasises that privacy could change (the user could, later, switch their account to Private, delete the tweet or their account altogether, but this wouldn't be reflected in the archive. We cannot know what is the user's original intent when they posted their message, and, as Paloque-Bergès (2010) noted, we aren't sure the user themselves know.

> "More often than not, in a technical ecosystem in which making content private is more difficult than sharing broadly, teens choose to share, even if doing so creates the impression that they have given up on privacy. It's not that every teen is desperate for widespread attention; plenty simply see no reason to take the effort to minimize the visibility of their photos and conversations." (Boyd, 2014).

Advantages over physical preservation

Mass preservation, easier access and organisation

Recording digitally material that's not born-digital

Cost of restoration; case studies of lost data

A number of institutions have already experienced loss of data, or data being corrupted. This can have disastrous consequences for businesses, for example, as noted by McAteer (1996, cited by Ross and Gow, 1999, p. iii): "of the 350 companies unexpectedly relocated by the [1993] World Trade Centre (NYC) bombing 150 [43%] ceased trading, many because they lost access to key business records held in electronic form".

Data can, obviously, be sometimes restored. Ross and Gow (1999, pp. 39–42, referenced in Harvey, 2012, p. 34) provide a number of case studies. Harvey (2012, p. 34) also gives the ironic example of the website for the "Functional Requirements for Evidence in Recordkeeping" project, directed by the University of Pittsburgh which was accidentally deleted; but thankfully was automatically saved by the Internet Archive.

There is still the possibility of restoring data by specialised carriers; however Pearson (2009, slide 15) notes that while this was generally effective, this did not work on some items. He also regrets that the

restoration loses all metadata in the process; this can be an issue for sensitive or classified information (for business or governmental archives) as the carrier would be able to see it; finally, it is a prohibitely expensive operation, and so a carefully planned preservation could be much cheaper (and obviously more reliable) than neglect. However, these services are always useful in certain cases. Harvey (2012, p. 41) cites the statistics by the data recovery giant Kroll Ontrack (2011): “hardware and system problems cause most customers to approach them (29 per cent), followed by human error (27 per cent). Software corruption or program problems, computer viruses and natural disasters are the other reasons listed.”

Defining paradigms for digital preservation

Pre-digital paradigms

Inability to preserve the medium, only the bitstream

Importance of metadata, relationship & context

Definition of terms

What to preserve / Selectivity

Harvey 5 pillars for selectivity (Harvey 2012, p. 26)

- ~~If **unique information** objects that are vulnerable and sensitive and therefore subject to risks can be preserved and protected~~
- ~~If preservation ensures long-term accessibility for researchers and the public;~~
- ~~If preservation fosters the accountability of governments and organisations~~
- ~~If there is an economic or societal advantage in re-using information~~

- If there is a legal requirement to keep it (NSF-DELOS Working Group on Digital Archiving and Preservation, 2003, p.3).

Banks reasons for domestic archiving

In *The Future of Looking Back*, Richard Banks (2011) analyses the reasons that lead us to keep or discard physical objects at home. He argues that while making a selection is a positive thing because it forces us to filter our belongings and focus on what is really meaningful, we also run the risk of ridding ourselves of things that we wish we would have kept in retrospect. He points out that the organisation of digital files make this trickier: because we we can keep pretty much everything on virtual space, it is easier to accidentally delete meaningful things alongside the less important. (Banks, 2011 p. 22).

In the context of personal archiving, Banks notes that in the past, the legacy we leave behind us was a highly selective one. Because our possessions are limited by physical space, our ancestors until now "have implicitly indicated that an item is significant through the simple act of keeping it. Although thousands of objects likely crossed their path over their lifetime, the constraints of money and space meant that only a few were retained, often the most significant." (Banks, 2011, p. 78). He does however argue that this is no longer true in a digital world, because new possibilities mean we simply accumulate digital objects without necessarily ordering them by importance.

Technical possibility of archiving

Pros/cons of archiving the entire internet

There are various controversies around whether we should automate the archiving of the Internet, or if we should operate a manual selection. UNESCO (2003, p. 73) notes that those who value a more comprehensive collection (i.e. archiving "all" the Internet) 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". This argument is particularly true because of the digital media: "unlike non-digital material such as paper-based artifacts, where there is a period of time in which to make selection decisions before deterioration of materials becomes an issue, the time frame for deciding whether or not to preserve digital materials is very short" (Harvey, 2012, p. 57). UNESCO (2003, p.71)

backs that claim: “it may not be possible to wait for evidence of enduring value to emerge before making selection decisions”.

Other people [Cook?] think that it would be preferable to keep a curated collection because of a higher technical quality, and this also allow us to settle [negotiate?] preservation rights with the content producers. Harvey adds that we also do not have the technical capabilities to find all the pages on the Internet, and even if we had we would not have the technical possibility to store the whole of the Internet (Harvey 2012, p. 58), again, this is also something recognised by UNESCO: “‘there are usually more things = more information, more records, more publications, more data = than we have the means to keep” (2003, p.70).

The UNESCO (2003) and Cook (2000) both however recognise that a middle ground should be found and that both approaches are valid depending on the goal of the preservation. For instance, the Internet Archive[IA] is going through an automated process (archive as many pages as possible) but, because of technical limitation, only preserves the text and dismisses all other images, stylesheets and external media, which could be critical in some instances. It seems evident that for valued documents these should be retained. The British Library’s Web Archive[ukwa] [http://www.webarchive.org.uk/ukwa/]

Dangers of selection (Cook)

However, Cook (2000) points out the dangers of manual selection for what is worth archiving: “With memory comes the inevitable privileging of certain records and records creators, and the marginalizing or silencing of others” (Cook, 2000, cited in Harvey p. 58). He asserts that the archivists must be impartial and objective, and ensure all points of view are kept to ensure our shared memory does not “becomes counterfeit, or at least transformed into forgery, manipulation, or imagination”. This is something that is clear with manual archiving, but could also be a problem, maybe unintentionally, with the automated discovery of content when archiving Web pages.

Automatic selection (Reputation model)

In order to keep the most valuable content, algorithms could be used. Kawano (2008) proposes to use reputation models, that attempts to

determine the relative value of a Web page by considering its "importance, fairness, trustiness, uniqueness and valuation". This is done using a variety of ways; the algorithm suggested by Kawano relies on the consistency of a web page, that is the amount of time it existed, the number of links it contains and the number of external links to this page (Kawano, 2008, p. 291).

Crawling

Manual curating

Special Collections for events (eg before elections, see webarchive.org.uk for other ideas)

Another interesting selection process that has emerged is the one of *specific archiving* when certain events are happening. For instance, the UK Web Archive[ukwa] has, on top of their topical-based collections (Energy, Health and Social Care...), a few event-based collections (Queen Jubilee, Olympic and Paralympic Games, London 2005 terrorist attacks...) which reflect the state of the Web at that given moment in history. Masanès (2006, p.83) notes that this type of archiving must be manually curated and thoroughly preparated. He gives a methodology for this, using presidential elections as an example: a time frame for the collection must be defined ("3 months before the election + 1 month after"), what type of sites will be archived and when ("the political parties' blogs and websites, every week; analysis, commentary and humourous/satire websites, every month; online newspaper articles, once").

Selection: by genre

There are different genres of websites: institutional website, blogs, personal pages, forums. [[Genre have been studied in the context of the Web to see how they replicate or diverge from the genre in the print- ing world (see Crowston and Williams 1997) - in Masanès p88]]. An interesting point about this is that often, robots (crawlers, or later data analysis) can automatically identify their type [Rehm 2002 -> check reference, Masanès:88] which permits the automatic generation of metadata, or simply to restrict a collection of web pages to certain Web genres.

Web archiving: what stands as

heritage on the web

Who should preserve

Digital media 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 "analog" 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 characterize the classic industrial model of print commodities". Self-publication (particularly on websites) is much more common, and this is also true for businesses and personal data. The Task Force on Archiving of Digital Information (1996, pp. 19–20) suggests that when, in general, "stakeholders disseminate, use, reuse, recreate and re-disseminate various kinds of digital information, they can easily, even inadvertently, destroy valuable information, corrupt the cultural record, [and ultimately thwart the pursuit of knowledge that is their common end]. Against such a danger, a safety net is needed to ensure that digital information objects with long-term cultural and intellectual value survive the expressions of stakeholder interest with their integrity intact.

[> Society has always created objects and records describing its activities, and it has consciously preserved them in a permanent way... Cultural institutions are recognised custodians of this collective memory: archives, librar[ies] and museums play a vital role in organizing, preserving and providing access to the cultural, intellectual and historical resources of society. They have established formal preservation programs for traditional materials and they understand how to safeguard both the contextual circumstances and the authenticity and integrity of the objects and information placed in their care... It is now evident that the computer has changed forever the way information is created, managed, archives and accessed, and that digital information is now an integral part of our cultural and intellectual heritage. However the institutions that have traditionally been responsible for preserving information now face major technical, organisational, resource, and legal challenges in taking on the preservation of digital holdings (B. Smith, 2002, pp.133–134) [cited in Harvey p. 29].]

A variety of new stakeholders are now required to collaborate more for a good preservation practice to take place. Lavoie and Dempsey (2004, cited in Harvey, 2012, p. 30) argues that cooperation "can enhance the productive capacity of a limited supply of digital preservation funds, by building shared resources, eliminating redundancies, and exploiting economies of scale". Chapter 11 of the UNESCO guidelines (2003) give more pratical details on how this can take place, giving structural models and tips on what to share (information, standards, division of labour...). [CLOCKSS cooperation to preserve journals (see Harvey p32 + chapter 9); + Harvey p32 National Library of Netherlands & Elsevier Science + most uni libraries access with publishers, even UoD with eg Dawsonera/ACM]

Legal issues

(see later) special case for emulation

In certain cases, the archiving process is not made in accordance with the original publisher. This is often the case for the automated archiving of the Internet (e.g. Wayback Machine), which automatically crawl pages, but as a website publisher, it is easily possible to opt out of these schemes [see http://archive.org/about/exclude.php [Accessed 17th Apr 2014]]. In the case of emulation, this is what happens with *reverse engineering*, that is, disassembling the original program to understand how it works and recreate a similar version which can operate from a different hardware or software architecture. This is a common process for proprietary software or hardware, whose source code is not available to the general public and no documentation is provided for developers. Often, this is developed by amateurs and hobbyists, but is illegal as it breaks the intellectual property of the original creator; Nintendo, for example, attacks reverse engineered emulators[http://www.nintendo.com/corp/legal.jsp [Accessed 17th Apr 2014]] and the sharing of ROMs [Read Only Memory; the bitstream of cartridges or discs containing the games]. There is however popular demand for playing these, and Nintendo has developed their own emulators to play legacy games on their new consoles (see the Virtual Console [http://en.wikipedia.org/wiki/Virtual_Console]), while still retaining their property rights (you need to buy the games again on an e-commerce platform). This however causes issue for the longer term preservation: there is no certainty that Nintendo will continue to provide this service on their future generations of consoles, or if they ever cease operations.

A similar issue happened with Apple Computer. In 2005, Apple changed the architecture (that is, the "language" of instructions needed for the processor) of their personal computers from PowerPC to Intel x86. This means that applications needed to be rewritten, sometimes from the ground up, to work natively on the newer Apple computers. To ease the transition, Apple created **Rosetta**, a software that dynamically translates PowerPC instructions to x86 instructions, so that buyers of the new Apple computers were still able to open legacy applications that were not yet rewritten for x86 — although this came with a significant performance hit. In 2009 however, support for PowerPC was entirely discontinued [...]

Libraries

Good curating and preservation starts early. Smith (2003, pp. 2–3, cited in Harvey, 2012, p. 32) suggests that "the critical dependency of preservation on good stewardship begins with the act of creation, and the creator has a decisive role in the longevity of the digital object". Harvey (2012, p. 32) notes that "for most creators of information in any form this is a new role"; therefore publishers of digital content are not necessarily well prepared for preservation, and this is why collaboration with librarians and archivists can be helpful (Ayre and Muir, 2004).

Ayre and Muir (2004) note that if the publisher worries about preservation themselves then they generally already own the rights for preservation, but there are negotiations to be done if an external body helps with preservation.

National libraries

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, Elsevier Science) or publishers directly (ACM). A report by Hedstrom and Montgomery (1998, cited by Ayre and Muir, 2004) found that most research librarians in the UK thought that a legal deposit to national archives was the best option, but this was not a

clear consensus. The report also highlighted that these librarians thought that “other libraries and publishers would need to, and should, have some involvement in preservation”.

Because all preservation methods involve copying of the original object, it might be illegal to make such a copy without the necessary rights, even if it is for preservation purposes. Certain countries have laws to overcome this and allow national libraries to preserve all materials; Ayre and Muir (2004) list a few of them and their specificities, but the situation has since evolved, as a lot of legislations have enacted changes to specifically include digital materials.

Archivists

New stakeholders

Profits vs nonprofits

Harvey (2012) points out doubts in the real interests of for-profit companies to achieve long-term preservation;

> Even given the prevailing market-driven political ethos, it is difficult to envisage a situation where market forces will be sufficient to ensure the preservation of this digital material. The opposite is more likely to apply: ‘in some cases market forces work against long-term preservation by locking customers into proprietary formats and systems’ (Workshop on Research Challenges in Digital Archiving and Long-term Preservation, 2003, pp.x-xi). (Harvey, 2012, p. 32). [link with Google concerns]

How to preserve

Software methods

~~Making hard copies~~

~~What appears to be the simplest solution is to make physical copies (i.e. print) of the digital documents. Although this certainly solves some~~

problems in regards to technological obsolescence, Rothenberg (1999) and Granger (2000) pointed out that this cannot be done for more interactive material besides text and images. There is also, obviously, an issue with the amount of content that can be stored in this way — this would not work for archiving websites. Granger (2000) recognises that "if one's sole concern is with the intellectual content of a document and the document is of a fairly simple nature [...] this at least provides some form of security".

Standardisation

Standardisation is the transformation of a document's content to a format that is standardised, open and well documented, ensuring that we will be able to open it with a variety of applications on a variety of platforms and that we can rebuild a program for visualising that document — as opposed to vendor-specific, closed source program that could be discontinued at any point in the future. Rothenberg (1999, p. 10) claims that this is a bad solution because "standardisation sows the seeds of its own destruction by encouraging vendors to implement non-standard features in order to secure market share" (therefore these non-standard features will be lost if opened in a different application). Granger (2000) argues that the blame is not to be put on the method, but on vendors. Standardisation is a type of migration and therefore share some of it cons (notably, this absence of vendor-specific features could imply a loss in the presentation and other characteristics). Standards eventually become obsolete so it is also paramount to preserve the specifications of the format, or to define formats that are unambiguous, such as CSV or XML.

A slightly different approach to this method is rebuilding viewers. In this scenario, the bitstream is not migrated to a standardised format, but kept as is, we simply re-build a viewer that is able to read that format. This is not as straightforward as it sounds, because most of the time, formats for proprietary software are not documented and decrypting them would require reverse engineering. If they are documented, then it is likely that this format is already a standard. There exists cases where reverse engineering is possible, or where a non-standard format is documented, for example, Microsoft Word files are not a standard, but there is plenty of documentation available online to produce code able to read them. An experiment by the VERS project (Thibodeau 2002, p. 22) has shown that it is possible to reengineer a viewer for the PDF file format from the specification available, as long as we also preserve these specifications, we

will be able to read PDF files in the future.

UVC

Lorie (2000, in Thibodeau, 2002, p. 22) has also proposed that the use of a portable programming language, such as Java [footnote: 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, 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.]. Lorie developed a software called Universal Virtual Computer (UVC), able to decrypt a variety of file formats, which can run on any platform supporting Java; because Java could be implemented on virtually any architecture, this means that a reader could be run on any platform without any effort. However, Lorie points out, this approach is limited in terms of functionality, and comes at the price of performance (because portable programming languages are platform-agnostic, this means the code is not optimised for any of the platforms). I am also wondering if this would really work on the long term, as we cannot be sure that future implementations of Java will be backward compatible with code written for a previous version.

Migration

Migration is the transformation of a document's content to a format that is not necessarily standardised, but at least current, or, the newer version of same format. This could be, for example, transferring documents from the WordPerfect 6 format (used in the 80s and late 90s) to a Word 2011 format, or a Word 98 file to Word 2011. This should be done regularly, as after a certain period of time new programs will stop having backward compatibility to import these old formats (Thibodeau, 2002). This method has several downsides, mainly being that it is labour intensive and expensive (Thibodeau, 2002; Rothenberg, 1999), product lines, and therefore the migration path, could be terminated at any point (Thibodeau, 2002), the presentation and other characteristics could be changed (although this can be overcome by migrating to more flexible formats like PDF or LaTeX). Rothenberg (1999) adds that this method is also prone to error, risky (as it could compromise the integrity of the data), not scalable and it requires new solutions for each new format. Granger (2000) argues however that

comparatively, other methods (particularly emulation) are potentially much more expensive and labour intensive, and that for most preservation programmes, migration remains the only viable method.

Computer museums

Another approach to preservation, proposed by Swade (1993), would be to preserve a variety of hardware, software and devices to access obsolete media in "computer museums". This is generally not considered a good solution; Granger (2000) notes that it would be unreasonably expensive to keep old machines running for an extended period of time, that computers chips will eventually decay anyway (and repair or replacement will be impossible), and that storage media not stand the test of time. There are still two possible use cases, identified by Rothenberg (1999): ensuring that emulators work as expected (by comparing them to original platforms), and helping with the recovery of lost data on obsolete media

Emulation

Emulation is the dynamic transformation of instruction sets to allow software and/or operating system designed for an obsolete architecture or platform to run on current systems. It is a highly complex method that offers certain benefits; notably, this is the most conservative method, as it is not destructive of the content or presentation. It allows us to run the original program used to create a digital object without altering that bitstream ; it "keeps the look and feel as well as the interactivity" (Granger, 2000) and as such, it is particularly interesting for objects rich in interaction, such as video games, data visualisations, or formats that are particularly complex (for example, medical imagery, 3D modelisations), for which methods like migration would be hard. It also ensures complete integrity and authenticity, which could be crucial for some data types.

There are numerous criticisms against emulation. A number of legal and intellectual property problems are raised, if the emulation strategy is not supported by the original author [see Nintendo & Apple]. The complexity involved in creating emulators might be not worth it and requires a coherent organisation to reduce overhead (Granger 2000). Additionally, Thibodeau (2002, p. 21) argues that keeping the original functionality is not necessarily a positive thing when it comes to the delivery of the preserved content; "most users in the future will never have encountered—not to

mention learned how to use—most of the products they will need to access the preserved information objects", and "it would cut users off from the possibility of using more advanced technologies for discovery, delivery, and analysis".

> "Emulators themselves become obsolete; therefore, it becomes necessary either to replace the old em- ulator with a new one or to create a new emulator that allows the old emulator to work on new platforms. In fact, if you get into an emula- tion strategy, you have bought into a migration strategy. Either strategy adds complexity over time." (thibodeau 2002 p.20)

[NB: due to the length limit of this literature review, I have only explained the most popular archiving techniques. See Thibodeau (2002) and Harvey (2012) for a more comprehensive list of methods, including configurable chips, persistent archives, object interchange format, etc.]

Choosing one

Thibodeau (2002, pp. 15–16) suggests four criteria to choose a preservation method: *feasibility*, *sustainability*, *practicality* and *appropriateness*.

Feasability means that the hardware and the software for implementing a given method must be existing and developed. Sustainability is ensuring that this method can be "applied indefinitely into the future, or that there are credible grounds for asserting that another path will offer a logical sequel to the method, should it cease being sustainable". This means that the method must be resistant to technological obsolescence, both for storing the data (medium will change) and accessing it (means to access the media ane decode the bitstream will change).

Practicality is establishing that this method will be reasonably difficult and affordable. [This can generally change depending on the resources of the preserving organisation, the available technology at that time and the valuation of the digital objects.]

Finally, appropriateness is using a method that is relevant to the type of material being preserved, and the objectives of the preservation. For example, del Pozo et al. (2012, p. 7) suggests that if the integrity and

authenticity of the documents are not paramount, then a migration that keeps the contents but discards the presentation might be the least expensive and simplest option. Inversely, when trying to preserve highly interactive material such as video games or when the specifications for a format are unknown, then emulation is a more pertinent option, because migration would be extremely difficult and inaccurate (Granger, 2000).

Authenticity/integrity and issues with each method

Different preservation methods yield different results in what is called **authenticity**, that is how faithful the rendition [its presentation] of the preserved object will be. Authenticity is subjective and therefore there is debate around this issue (Del Pozo et al., 2012, p. 7; Cook, 2000); however most agree that while as little difference in the presentation is desirable, it is rarely (if ever) possible to keep a truthful rendition. As an example, del Pozo et al. (2012, p. 8) use a spreadsheet within the migration method:

> "For instance, one institution may decide to normalise a spreadsheet into a PDF, which would retain the cell values but destroy any formula used to calculate them. This would favour a particular presentation of that digital object over the information form of the original. On the other hand, another institution might decide to normalise spreadsheets into ODF, which may lose some formatting but retain the formula used to derive each cell. In this case, being able to retain the information form of the file would be seen as more important."

Thibodeau (2002, p. 28) however points out that "people will want to use the best available technology—or at least technologies they know how to use—for discovery, retrieval, processing, and delivery of preserved information. There is a danger that to the degree that preservation solutions keep things unaltered they will create barriers to satisfying this basic user requirement." An overly faithful preservation would therefore not be desirable. Again, this is subject to a high number of parameters.

Another important issue is *integrity*, i.e. "ensuring the object is complete and has not been altered or corrupted" (Harvey, 2012, p. 54; Ross [CITED], 2002, p. 7), that is, the object "is what it purports to be", and no parts are missing from it. This is particularly important, Harvey (2012, p. 54) notes, for business archiving, where the legal value of a digital object could be lost if we cannot demonstrate that this object hasn't been tampered with, or that a

migration processes have not incurred data losses. However, as Lynch (2000) has argued, the integrity of objects is only a matter of trust. The guidelines of UNESCO (2003, p. 22) recommend that the best protection for authenticity is proper documentation of the archival process; Harvey (2012, p. 54) expands on this idea: "for instance, if it can be established that a digital object has always been kept in the custody of an archive in which every change to it has been recorded (such as changes resulting from migration of the bit-stream – the dates at which migration occurred, to which media, and so on), then we can be more secure about its integrity."

Emulation / legal rights with videogames

Use of standards

Some standards such as OAIS (Open Archival Information System), originally developed for space science, could be used. Sadly [no implementation][source??]

Personal archiving

On a more personal level, Harvey (2012) and Lukesh (1999) note that personal correspondance might be lost if we do not take steps to preserve them.

> "The widespread shift from writing letters to the use of email has diminished the likelihood that personal correspondence will remain accessible to future historians. Lukesh asked in 1999, 'Where will our understandings of today and, more critically, the next century be, if this rich source of information is no longer available?', as scientists, scholars, historians and almost everyone increasingly use email" (Lukesh, 1999) (in Harvey, 2012 p. 32)

While this is a correct idea, I think that, just like the archiving of Usenet, 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 artifacts 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, redundance) (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; second, storing photos and data on online services mean we may forget about them in a few years (Banks, 2011).

What happens after our death

[Digital heritage]

Shifts in usages (more photos)

Advantages (sharing, organisation etc.); record part of our lives we never did before

Digital artifacts have certain fundamental advantages over physical objects. the main two being 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). This actually provides numerous advantages in terms of preservation.

[[+ tie with "In the past ancestors have signified something is important simply by keeping it" part]]

The first one, ease of organisation, can be significantly helpful in tackling the abundance of digital data. Banks (2011) suggests that the technologies lying behind the ranking in the result of Web search engines could, eventually, help us in understanding and finding what is our most important data. An attempt has been made by the MUSE (Memories USing

Email) project (Hangal et al., 2011), which aims to find patterns in our emails 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 thereby use their email account as an informal backup device. Therefore, email archives contain or reflect memories that are extremely valuable for the purposes of reminiscence." (Hangal et al., 2011)

Early users reported that they have been using it to: make a summary of their work progress over the year; 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's been left unfinished and forgotten; renewing with old relationships; and serendipitous discovery. The authors note in the conclusion that the different uses of their program went beyond their 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."

The second advantage, the ability to accurately and easily make copies and backups, is fairly obviously an advantage for personal preservation. However, Banks (2011) notes, the process of backing up is troublesome and storage media do not make our data immune. Additionally, the use of online services to store our data creates an abstraction between the storage material and us; about Flickr: "I have no idea what kind of hardware my files are now stored 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).

Keeping memories attached to digital objects

Banks (2011) also suggests that we could blend in the physical and the digital to allow recording of things we never had the opportunity to before. He gives as example the Weather Camera, an experimental object by Kjell Wilkens (2011) that records wind and temperature along with commentary,

instead of an image, as an alternative to recall a place; or 3D backups by Héctor Serrano (2009), another experimental project that allows us to make backups of sentimentally important objects through 3D scanning and 3D printing. Similarly, Banks notes tools by Microsoft Research to recreate places in a 3D environment based on photos, allowing to recreate a sentimentally important place at any given point in time provided enough photos are available. A similar technique is used in bigger preservation projects, as described by Zhou et al. (2012), to make high quality 3D models of culturally important places and objects in China.

Kirk and Sellen (2010) found however that ultimately, stories and narrative are more important than the preservation of the artifact itself. They propose that "technology might play a role in capturing and associating stories or narratives with different physical objects", imagining that we could augment physical objects with RFID tags (and I would add, virtual reality technologies) to pass down stories attached to these objects.

Conclusion