The need for digital preservation

With an exponential growth in data being produced digitally (SINTEF, 2013), comes a growth in the volume of data worth preserving 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 artifacts, "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 artifacts and our personal heritage is subject to discussion.

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. For example, the Internet Archive (Kahle, 1996) stores copies of as many websites as possible, but discards all images, stylesheets and multimedia objects in the process due to technical limitations; evidently this is not appropriate for valuable websites. The British Library's Web Archive [footnote: http://www.webarchive.org.uk/ukwa/] curates collections of websites by topic and by events (Queen Jubilee, Olympic and Paralympic Games, London 2005 terrorist attacks, General Elections ...) 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 "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).

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; for example D3 cartridges tapes start to deteriorate after 50 years in optimal conditions (low temperature and humidity) but only a year in warm and humid atmospheres. The range is even greater with optical media like DVDs, going from 3 months to a theoretical 200 years.

But the media are only part of the equation; because these media 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 software is left.

Who must preserve, and what we should keep

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 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 have only explained the most popular archiving techniques. See Thibodeau (2002) and Harvey (2012) for a comprehensive list of methods, including configurable chips, persistent archives, object interchange format, etc.

Making hard copies

What appears to be the simplest solution is to make physical copies of the digital documents (i.e. print them). Rothenberg (1999) and Granger (2000) both note that this obviously isn't a viable option for large data sets and interactive content, but Granger recognises it could provide a form of security for simple documents.

Computer Museums

Swade (1993) proposed to preserve a variety of hardware and software to access legacy media and files in centralised places. Granger (2000) critiques this idea, thinking that it would be prohibitely expensive and would only postpone the issue, as computer chips will decay anyway. Rothenberg (1999) has the same arguments but noted that such museums could be helpful for recovering data found on obsolete media, and testing emulators.

Migration

The most popular option is migration, that is, the transformation of a document's content to a current format — either a different format of the same type (e.g. image type, text type) or a newer version of the 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. 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.

Standardisation

This option is similar to migration, but it preconises switching to a software-agnostic standardised format, which must be well documented or unambiguous, to allow the re-creation of editors and viewers later (Granger 2000). This solution avoids some of the drawbacks with migration (it avoids the need to migrate periodically and the format/software cannot be discontinued) but, because no vendor-specific formatting is allowed, it could lose some specificities of the document and as such is only usable for certain simpler data types (Rothenberg 1999).

Universal Virtual Computer (UVC)

This model proposed by Lorie (2000, in Thibodeau, 2002, p. 22) uses a portable programming language [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 (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.] to create an implementation of a viewer which could then theoretically be ran on any later platform supporting that programming language. However, Lorie notes that this approach restricts functionality

and performance, and I am wondering if future implementations of these portable programming languages will be backward compatible with that code.

Emulation

Emulation is the most conservative model; it is not destructive of the original bitstream (unlike migration) and proposes to emulate the characteristics of a legacy platform on a current computer, in order to run the original software used to create it. Therefore it "keeps the look and feel as well as the interactivity" (Granger, 2000) and is particularly useful for highly interactive material like video games and scientific visualisations. It also guarantes integrity and authenticity, which could be crucial in some preservation scenarios.

However, it is highly complex and is criticised for this; Granger (2000) thinks that the amount of work involved in creating emulators would rarely be justified, and if it is, it requires a coherent global organisation to reduce overhead. Emulators also become obsolete (Thibodeau, 2002, p. 20) and so still require a high maintenance cost. Additionally, Thibodeau (2002, p. 21) argues that keeping the original functionality is not necessarily a positive thing when it comes to delivery: it would deprive new users of future technological advances for discovery and analysis, and force them into learning to use a dated interface they might never have seen before.

It might, also, raise a number of legal and intellectual property issues from the software provider. For software that is proprietary or discontinued, the source code is not available, and being able to emulate it requires to do reverse engineering [footnote: Disassembling the original program to understand how it works and recreate a similar version which can operate from a different hardware or software architecture], which breaks intellectual property of the publishers.

Emulation is highly used for video games, as other methods are unsuitable for accurately recreating them. 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 or discontinued and so are the games. There has been a thriving community of hobbyists reverse engineering console games; this requires considerable efforts (byuu, 2011)

and is generally considered illegal. In response to the growing interest, and to protect their rights and interests[footnote: See http://www.nintendo.com/corp/legal.jsp [Accessed 17th Apr 2014]], Nintendo started to offer their own emulation solution to play classic games on their current generation consoles. 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. This has happened in the past with Apple, who provided emulation software to run legacy software on their new architectures to ease transition (MacInsider, 2011; Mesa, A. F., 1997) but was later discontinued, effectively removing support for any pre–2005 application on their current models.

Choosing a method

Thibodeau (2002, pp. 15–16) suggests four criteria to choose a preservation method: * **Feasibility**: means that the hardware and the software for implementing a given method must be existing and developed. * **Sustainability**: ensuring that the method is resistant to technological obsolescence and that it 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" (Thibodeau, 2002). * **Practicality**: establishing that the method is reasonably easy and affordable, in line with the preserving organisation's resources. * **Appropriateness**: the method must be relevant to the type of material being preserved, and the objectives of the preservation. For instance, del Pozo et al. (2012, p. 7) suggest that if the integrity and authenticity of the documents are not paramount, then a migration that keeps the contents but discards its 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).

Web Archiving

Archiving the Internet and preserving privacy

Camille Paloque-Bergès (2011; 2013) suggests 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 protocol 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 a 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 we know it, as it did not exist and is a separate network (albeit functioning on the same Internet). A number of privacy concerns were caused with the publication of searchable archives (first by DejaVu News in 1995, then Google Groups in 2001); messages who were, 15 to 20 years earlier, thought to be confidential are now considered as our common digital heritage.

Similar concerns were raised at the announcement of the archival of Twitter by the Library of Congress in 2010, but two significant elements make it different: the archive is not public and is made by a not-for-profit body. On the mailing list of the Association of Internet Researchers (2010), 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) [footnote: Direct link: 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, 2010 [footnote: Direct link: http://listserv.aoir.org/pipermail/air-l-aoir.org/2010-April/021136.html]; see also his paper (2010) on Facebook privacy, expanding on that argument).

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 corespondance and sentimental artifacts stored digitally might be lost if we do not take steps to preserve them, and wonder how we will understand our modern life without these exchanges. 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, 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 artifacts 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's 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 artifact 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).