Evaluation and Strategies of Digital Preservation & UNESCO's Role in Facing the Technical Challenges

Margriet van Gorsel, Michiel Leenaars, Natasa Milic-Frayling, Jonas Palm

Introduction

PERSIST (*Platform to Enhance the Sustainability of the Information Society Trans globally*) is a collaborative project of UNESCO, IFLA, ICA and other partners to address globally pressing questions on the preservation of digital information in the public domain related to strategies, technology, selection issues, and roles and responsibilities. It works under the assumption that on these issues a high-level global policy discussion is needed between heritage institutions, IT-industry and government, and that UNESCO's Memory of the World Program is a unique platform to conduct that discussion.

The idea for PERSIST was born at the Conference *The Memory of the World in the Digital Age: Digitization and Preservation* in Vancouver (September 2012). The Declaration adopted by its participants states that

> *there is a pressing need to establish a roadmap proposing solutions, agreements and policies, that ensure long term access and trustworthy preservation. This roadmap should address issues like open government, open data, open access and electronic government. It should dovetail with national and international priorities and be in full agreement with human rights.*

PERSIST was launched as a project at the Conference *A Digital Roadmap for Long-Term Access to Digital Heritage* in The Hague in December 2013. In its initial stage the project is coordinated by the Netherlands National Commission for UNESCO and financed by the Netherlands Ministry for Education, Culture and Science.

The work for PERSIST is divided into three task forces: content, technology and policy. For the technology taskforce 'preservation strategies' has been chosen as the first subject of attention. Responsibility for the work in the technology task force is divided between Margriet van Gorsel (National Archives of the Netherlands), Natasa Milic-Frayling (Microsoft Research), Jeanine Tieleman (DEN), Michiel Leenaars (Internet Society Netherlands), Jonas Palm (Memory of the World Sub-Committee on Technology) and Vincent Wintermans (Netherlands National Commission for UNESCO).

Synthesis

The start of this position paper lies in a discussion in spring 2014 in Amsterdam. This discussion fed the idea of a global service for enhancing digital preservation practices by jointly addressing the issue of sustainable computation. At the ICA 2014 conference we will host a workshop to reflect on this idea. Three persons agreed to share their thoughts on this subject and contribute to this position paper, which is intended to set the stage for a discussion at the ICA-workshop. They look at the problem of the sustainable information society from different perspectives. Jonas Palm contributed a *Reflection on the long standing approaches of memory institutions* to use technology and techniques to establish continuity of content. Natasa Milic-Frayling takes us to the *Computing Ecosystem and our Digital Legacy* and finally Michiel Leenaars gives us a look *Beyond a software and hardware repository*.

All three writers reflect on the purpose of libraries and archives to ensure access to information. The knowledge base of their profession lies in appraisal and selection, collection management and services to the general public. Some institutions have acquired new specialisms, for example preservation and conservation of physical information carriers but mostly they outsourced those special tasks. The budget for managing their collections comes mainly from the government's expenses for cultural, educational and scientific activities. In the pre-digital era libraries and archives had time on their side to spread the cost of managing and maintaining their collections.

And then digital information became the standard. And librarians, archivists and keepers of digital heritage had to acquire a whole gamut of new skills to acquire a level of certainty about the sustainable access. Earlier on only keepers of audio-visual collections were facing that kind of difficulties; now every heritage institution is beset by never-ending and always pressing technical challenges. In the position papers of Palm, Milic-Frayling and Leenaars we read what changes occur in the profession of heritage institutions since the introduction of digital information.

To give these papers some background it's good to know that some common understanding of the difficulties of digital preservation has been achieved in the past decade. Colin Webb, National Library of Australia, wrote in *The Memory of the World: Guidelines for the preservation of digital heritage*[1]

> *Digital preservation consists of the processes aimed at ensuring the continued accessibility of digital materials. To do this involves finding ways to re-present what was originally presented to users by a combination of software and hardware tools acting on data. To achieve this requires digital objects to be understood and managed at four levels: as physical phenomena; as logical encodings; as conceptual objects that have meaning to humans; and as sets of essential elements that must be preserved in order to offer future users the essence of the object.*

The Guidelines then elaborate on the dynamics of preservation processes, that almost always involve making changes – transferring data from one system to another, from one carrier to another, adding or updating metadata, creating new copies that need new file names, changing the means of presentation as technologies change, and so on. The

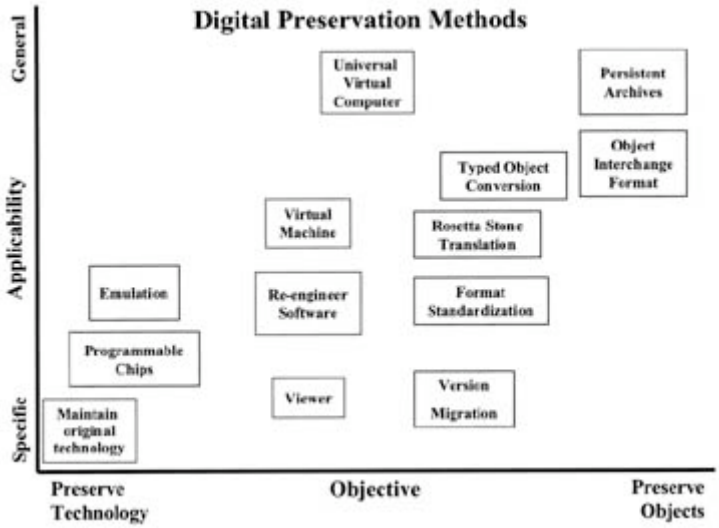
[1] The Memory of the World: Guidelines for the preservation of digital heritage (2003) page 34

Guidelines then address the issue of the safeguarding of the authenticity that derives from being able to trust both the *identity* of an object – that it is what it says it is, and has not been confused with some other object – and the *integrity* of the object – that it has not been changed in ways that change its meaning.[2] This implies that there is always a dependency between data and software: all data require some kind of software in order to be presented in an understandable form to a user.[3]

To quote Kenneth Thibodeau in his 2002 article *Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years1*[4]

> *One can only preserve the ability to reproduce the document.*

He had already made this graph to illustrate the spectrum of digital preservation methods.



Thibodeau continues by saying that solutions for sustainable access of digital content have to apply to four criteria

- *Feasibility requires hardware and software capable of implementing the method […]*
- *Sustainability means either that the 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[…]*
- *Practicality requires that implementation be within reasonable limits of difficulty and expense[…]*
- *Appropriateness depends on the types of objects to be preserved and on the specific objectives of preservation[…]*

He ends his article by saying

[2] Item page 109
[3] Item page 120
[4] Kenneth Thibodeau "Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years1" http://www.clir.org/pubs/reports/pub107/thibodeau.html

> *Finally, there is an overriding need to select and implement preservation methods in an open-ended system capable of evolving in response to changing needs and demands.*

So it's logical that one of the first initiatives of libraries and archives focused on the need to keep information about formats, after the software that creates them goes out of use. Examples are the Software Preservation Society, http://softpres.org, for the game industry and the Software Sustainability Institute, www.software.ac.uk/what-do-we-do/preserving-software, for researchers. Moreover they invested in the development of tools and registries that identify the file formats, show how it works, what it tells about the information what the information is exactly and, if necessary, how to transform from one format to another in all its consequences. These developments, like Pronom, The National Archives UK, and Totem, University of Portsmouth, are very important because we learn about the influence of hard- en software on information. The lesson learned is that the more standardized and open we work the less we have to do at the back end of preservation.

And that is where this position paper, written 10 years and much experiences later comes in. Leenaars, Milic-Frayling and Palm all give their interpretation of the challenges of digital preservation. Palm emphasizes the critical need for memory institutions to keep searching for new ways to do their age-old job of preserving their information. In origin they look at back end solutions of digital preservation but already memory institutions started to look at new business models for instance by mobilizing the crowd, by influencing governmental records management and by supplying a trusted digital repository. They acquire, store, preserve - not only against decline of the information carrier but also to keep it readable, integer, trustworthy, authentic - and give access. Unfortunately time can't help to spread the cost of preservation because obsolescence lies around the corner and they need more specialisms in- or outsourced. That preservation is about strategy and co-operation, has become increasingly acute since the appearance of digital heritage. It is from this perspective that the PERSIST project aims to give special attention to working with the IT-industry for front-end solutions for digital preservation.

Palm is backed up by the publication *From Theory tot Action: "Good enough" Digital Preservation Solutions for Under-Resourced Cultural Heritage Institutions*, a Digital POWRR White Paper for the Institute of Museum and Library Services[5], that states that small memory institutions have

> *a lack of available financial resources; limited or nonexistent dedicated staff time for digital preservation activities; and inadequate levels of appropriate technical expertise. Some of the case studies also mentioned a lack of institutional awareness of the fragility of digital content and a lack of cohesive policies and practices across departments as a contributing factor towards the absence of real progress.*

[5] *From Theory tot Action: "Good enough" Digital Preservation Solutions for Under-Resourced Cultural Heritage Institutions*, a Digital POWRR White Paper for the Institute of Museum and Library Services, J.Schumacher, L.M.Thomas, D.VandeCreek a.o (2014). page 4
http://commons.lib.niu.edu/handle/10843/13610

The up side is that there are things to do even if you are a small institution. The article gives four pointers[6]

1. *Understand that digital preservation is an incremental process. Digital preservation is achieved through cumulative activities of increasing efficacy. It is time to embrace a "good enough" approach to digital preservation.*
2. *Focus on a set of discrete activities that can immediately yield higher levels of preservation, however modest. These can include inventorying your existing content, educating content creators, and designing an ingest workflow.*
3. *Examine your institution's strengths and potential challenges to committing resources to its digital legacy. Understanding where you are, where you want to be, identifying roadblocks preventing you from getting there, and connecting with allies who can help you move further along the way is crucial to this process.*
4. *When exploring more robust technical solutions, understand that selecting more than one tool or service may be preferable.*

That brings us to the question what else there can be done looking outside the heritage-world, further on the axis of Thibodeau in the direction of cooperation between the world of technology, heritage institutions and government? Are there different business models and ways of opening up the interdependence between the public and private world by looking at front end solutions?

Milic-Frayling gives a very tempting expose on the connection and possibilities for strengthening the business models of memory institutions and IT-partners. Innovation, continuity and profit are key in the commercial world driven by supply and demand. The public world of government and cultural heritage understands those critical indicators but have a different responsibility namely the equal accessibility of means in society, supporting accountability and compliance of government and solidarity between communities. And besides a lot of innovation is supply driven. Milic-Frayling explains that these different business models can be brought together in three ways:

- Government and memory institutions must develop to a 'significant player' that can negotiate their interest in the world of industry. They have to keep their eyes and minds open for developments like Big Data because in that world the reuse of digital information will be key. It stimulates long term storing and accessibility.
- Bring in the Mediator to find the common ground between the technical provider and the heritage institution.
- Make preservation a part of design and engineering and make provisions to be backward compatible by developing a computing ecosystem.

All three of her propositions have potential and looking from a UNESCO perspective are globally available. A critical point must be made on the backward compatibility by making an ecosystem because the management of such a service will be difficult, costly and raises questions about the responsibility.[7]

The paper of Leenaars addresses digital preservation from a different angle. He is the one that explores the possibility of a front end solution by looking at the interoperability-over-time. He explains what it means when digital information is presented to the user

[6] Item page 15

[7] Added an updated text from Milic-Frayling just before the ICA conference.

on different machines or in different software surroundings and what happens behind the screen. From that point of view his expose presents research on provenance, dynamics and transfers of information and the consequences for those important keys for memory institutions namely authenticity, completeness and trust.

He relates the ideal digital world to reality and makes an urgent appeal for 'a healthy group equilibrium' between all the stakeholders both public and private. He gives several suggestions to move forward:

- Getting to know and trust each other, if public and private partners want to see what's happening behind the screen relevant applications must be mutually available for testing.
- Get developers from the 'open society' involved in an early stage.

Leenaars focuses on the relationship between the trustworthiness and the 'traces' documents contain of production, presentation and use. In view of the enormous volume of documents and the variety of formats as websites, databases, audiovisuals, compound information for instance geo-data and on-line-forms, the question remains if these formats can get that same attention or that other solutions will be necessary. The propositions of Leenaars enhance the quality of digital information in such a way that looking from a UNESCO perspective transparency and interoperability are enhanced and in the long run digital preservation will be less costly, globally reachable and manageable. The downside is that it is still in a developments phase and doesn't bring immediate relieve, it will be costly and also raises questions about the responsibility.

Linking Pin

UNESCO is uniquely positioned to function as the focal point for this ambitious undertaking. With its 195 Member States its outreach is nearly universal. Its standard setting work for cultural heritage is widely acknowledged. In the field of documentary heritage it has adopted the Charter for the Preservation of the Digital Heritage (2003) and is currently elaborating a similar instrument for documentary heritage in general. It has as a specialized agency of the United Nations easy access to governments, but it also maintains relations with a large network of NGOs. ICA, and similar organizations in the field of heritage like IFLA, have an official associated status with UNESCO.
Documentary heritage has been in the orbit of the organization since its very beginning. UNESCO's Constitution speaks about 'assuring the conservation and protection of the world's inheritance of books, works of art and monuments of history and science' and 'initiating methods of international cooperation calculated to give the people of all countries access to the printed and published materials produced by any of them'. (Article I, 2 (c)) UNESCO's work for documentary heritage is carried out through the Memory of the World (MoW) Program. The MoW Register, that currently lists some 300 important collections and items from more than 100 countries, is the most important means by which UNESCO informs policy makers and the general public about the importance and beauty of this part of the world's heritage.

All writers agree on the critical role of the interdependency of the used software or as Leenaars calls it *the non-default content elements* or in the words of Milic-Frayling the *user experience*. From the three approaches presented both Palm and Milic-Frayling are back end solutions and the approach of Leenaars is a front end solution aiming at interoperability-over-time and strengthens the link between content and technology.

Memory institutions will have a hard time to handle this on their own so it's time to make the connection with the other players in the field on a technology and business level.

To start the discussion

The sustainable information society will profit from better public private partnership. So it's wise to consider the feasibility of 'future proof' business models for instance to search for new partners that also are concerned with the continuity of content as Big Data service providers are now. From a technology point of view we can ask ourselves:

- Is the creation of a global repository of software for preservation purposes (ecosystem) a good addition on the current preservation strategies?
- What do we think of the idea to organize worldwide automated interoperability testing? Can we make the problems and solutions visible in such a way that it leads to a global interoperability approach of policies and standards?
- Are the proposed strategies competitive and mutually exclusive or are they supplementary?
- How would the proposed repository overlap or complement existing initiatives like PRONOM and Totem?
- Is there an in-between?
- Are the proposed strategies universal?

And finally:

- in which way can UNESCO support these initiatives and thereby support the sustainable information society?

1. A reflection on the long standing approaches of memory institutions

by Jonas Palm
Head of Preservation at *Riksarkivet*, Sweden - President of UNESCO's Memory of the World Sub-Committee on Technology

We are in the middle of a paradigm shift in preserving documents: we are stuck in it and we have not yet found the solutions we need and - will we?

Until the beginning of this paradigm shift, preserving information was an issue of using or producing carriers with long-term stability: palm leaves, papyrus, parchment and paper with inks and colours that could stay inert as long as possible. When photographic processes were invented long-term stability was an early issue.

When documents were damaged they were mended or copied. Even when photocopying and microfilming became widely used during the 20th century for copying and securing information on deteriorating carriers, aspects on long-term preservation of these copies were important.

The easiest way to preserve these documents or copies has been to store them as properly as was possible, keeping them away from humidity and water, heat and fire. After being put on a shelf, they could be left for centuries and the information would be retrievable at any given moment.

Thus we have a long continuity of using and preserving man-readable documents. Documents where information is retrievable without any technology, though in the case of microfilm a magnifier is needed of course.

When audio recording was introduced, the paradigm shift into the era of machine-readable documents started. Recorded information was retrieved with intermediate technology. Technology developed new ways of recording sound and images during the 20th century, but it was at a pace which the preservation community could handle.

So far, the production of media with long-term storage properties was a viable business option.

When digital technology developed a completely new situation evolved. Media had until then been used to secure and distribute information of all kind with the aim of being more or less permanent. Digital technology was primarily a tool used to process, analyze and distribute information. Thus the aspect of long-term preservation of the information was not an issue for producers of digital technology and still is not.

To make things even worse for preservation, both hardware technology and software development is so fast and changes so much that older media- and file- formats and technologies become obsolete within a decade.

The simple way to preserve documents is not suitable for machine-readable documents. They cannot be left on a shelf as man-readable can.

Machine-readable information must be copied or merged to recent media to be secured for the future and it is a race against time. Old equipment is quickly disappearing and so are spare parts. Old file formats eventually can no longer be read by even the same program producing the file originally. Hardware can only read or write so many generations back or forward etc.

There is no sense in trying to find media which will last for centuries any more. The preservation issue is no longer about the longevity of the medium or the storage conditions. It is now about strategy and co-operation.

Strategies must be developed in co-operation with the industry and must be aligned continuously with developments in the industry, just as traditional road maps (meaning those we use when we travel) are updated from time to time to keep pace with reality. Technical development is quicker than our efforts to deal with it and everyone will use anything that is out there.

Two major roads emerge ahead of us – the migration of the digital information from generation to generation of hardware so that new generations of software can deal with it and present it as it was once created.

And/or, we are steadily building up a backlog of document formats, which can no longer be interpreted or will become inaccessible in a very short time. There must be strategies how to preserve these or access to the information the best way.

Along these efforts we will still have the old traditional man-readable documents properly stored and luckily this is much cheaper in the long run than long term storage of digital information.

2. Computing Ecosystem and Digital Legacy

Natasa Milic-Frayling, Principal Researcher, Microsoft Research Cambridge, UK

The involvement of ICT companies in a global approach to document and software archive as a contribution to continuity of content and focusing on a practical application.

Executive Summary

Digital obsolescence as an economic effect. Obsolescence of digital technologies, including hardware, software, and services, is a result of the continuous innovation in computing that is fuelled by prolific use of Information and Communication Technologies (ICT) across industries and other sectors. A digital technology becomes obsolete when it cannot be economically sustained because of the changes in the market place, such as lessened demand, that make its production and maintenance unfeasible. Thus, an attempt to arrest digital obsolescence is an intervention that defies market forces. It essentially prolongs the 'life' of technology beyond the point of sustainability in the original market. Such an effort could be supported for a while by funding agencies or volunteering work. In practice, these models are often applied to services that provide intangible value to the society and cannot be directly translated into an immediate revenue stream. Another approach requires solutions with sufficient demand and adequate business models to cover the costs of sustaining preservation services.

Fundamental dependence on computing technologies. As part of the preservation process, the archivist focus on the selection, curation, maintenance, and access to preserved artefacts, physical or digital. Maintaining the integrity of such artefacts is an important part of preservation and typically involves third party services such as sensor instrumented storage, shelving and encasing of rare manuscripts, video and audio tapes, and similar. Digital artefacts, including documents, databases, Web sites, and games, cannot be used without appropriate software and supporting hardware. Thus, maintaining the integrity of digital artefacts requires addressing the availability of software and storage of files, and any service that stores and instantiates legacy digital artefacts will have an ongoing dependence on the computing industry.

Redefining the business engagements. Creating economically viable preservation operations requires building a business relationship with the providers of software and hardware that became obsolete and negotiating the terms of use of their legacy products. Use of software and hardware in the peak of their market value is governed by licenses and contractual agreements that support the business model of the technology providers and ensure sales and economic viability. When technology transitions to the long tail, with a sporadic use and low demand, it cannot be further developed and maintained within the same business framework. Any service that provides access to the technology at that point may find it necessary to revisit the contractual agreements and renegotiate the terms of use in order to enable an alternative business framework.

Digital preservation solutions are digital. Maintaining the integrity of digital artefacts for future use involves aspects that are analogues to the preservation of physical artefacts, such as storage, retrieval, and curation. In the case of digital media, all three are tied to the computing technologies, including: (1) hardware and software to store data, content, and program files, (2) hardware and software to instantiate digital artefacts, and (3) hardware and software to support preservation operations such as risk assessment and planning, metadata generation, and customer facing services. All of these technologies are produced within the *contemporary computing ecosystem* and optimized to meet the needs and stay economically viable.

Status of digital preservation solutions. Technologies for storing digital files have been relatively stable and interoperable and their cost has exhibited a favourable trend, providing more storage at lower prices. Technologies for accessing digital objects are

available but only some have been explored and optimized. Current preservation practices involve *file format migration, hardware and software emulation*, and *software virtualization*.

The migration of file formats leverages the contemporary software to gain access to the past content. In some migration practices both the original file and the original software are abandoned, reducing the cost of file storage. However, the migration operation intrinsically grows at the rate of content production. Applying it perpetually involves accumulation of content that needs to be moved to the next generation software. The file migration approach is not applicable to digital artefacts that are highly interactive such as simulations and games.

The cost of emulation and virtualization is primarily related to the development and maintenance of emulation and virtualization software. New virtual machines (VMs) may need to be developed as the underlying computing architecture evolves. Any new software application eventually needs to be virtualized. While the core cost of emulation and virtualization does not necessarily decrease, it is independent from the rate of content production. The number of VMs that need to be provided depends on the demand to access digital content which could be exploited for revenue generation.

Leveraging the trends in the contemporary ecosystem. The key to economic sustainability of preservation initiatives is to leverage the trends in the computing ecosystem and become a significant player in shaping the demand. Fortunately, the emerging cloud computing platforms provide services for virtualization of computing environments and storage of large content repositories that are well aligned with the digital preservation needs. Furthermore, large volumes of digital assets that typically present a challenge to Memory Institutions are emerging as an asset in the global market, the 'Big Data'. Industries are looking for ways to leverage historical data to fuel innovation in specific domains. The trend towards 'Big Data analytics' continues to increase the demand for the cloud computing infrastructure, tools, and services and the cost and the quality are being optimized. Data processing scenarios that are part of the preservation services can therefore be supported efficiently by building on the main stream solutions. Furthermore, with clearly articulated requirements and a large enough constituency, the Memory Institutions can become a market segment that creates sufficient demand to dictate the requirements and affect the cost of such solutions.

Partnering with the ICT industry. In order to lower the barrier for creating sustainable preservation solutions, it would be beneficial to establish organizations that can mediate the relationship between the content owners and the technology providers. Such entities would serve as catalysts for cross industry engagements to address key issues, including software licensing agreements to accommodate new modes of technology use, access to software development tools and source code to enable porting of software to new computing environments, and safeguarding documentation about the past computing systems. They would provide a springboard for preservation solutions and insurance to content owners that their interests are represented. UNESCO is in a unique position to serve as a voice of the Memory Institutions and the societies, to foster partnerships with the ICT industry, to support development of standards for preservation practices, and to become a custodian of hosted legacy systems, tools, and documentation that are necessary to implement the preservation strategies.

Revisiting engineering practices. Looking ahead, requirements for digital preservation need to become a part of the design and engineering practices in the computing industries. Current focus of ICT industry is on productivity and reliability of products and services. Each enterprise is driven by opportunities to improve their technology during its market life-span. However, digital assets need to be exploited in the far future and the engineering practices need to consider the complete life-cycle of technology. That, for example, requires us to consider software architecture designs that allow easy

virtualization and porting of software to make the preservation of produced digital assets technologically feasible and economically affordable.

Digital Revolution and Digital Obsolescence

Digital revolution has been carried out by a broad and intricately connected ecosystem of enterprises that provide computing technologies, from broadly used office productivity tools to computing infrastructure that enables digital communication, data processing, and content delivery. These enterprises range from large corporations, such as IBM, Intel, Microsoft, Oracle, and Apple, to small and medium size enterprises that provide specialized solutions, and a wide range of non-for profit and open source initiatives that depend on community engagements. Together with a range of supporting services that enable distribution and adoption of computing technologies, they comprise the *contemporary computing ecosystem*.

This complex ecosystem is based on the needs and demands of the technology users and shaped by the business models that sustain the production of technologies. Moreover, each hardware product and a software application is heavily dependent on multiple technology providers. Even a simple personal computing (PC) device involves a large number of components, each provided by a specialized segment of the ICT industry. Besides the natural interconnection of these industry segments, each of them continuous to innovate in their area of speciality. As a result, the users are provided with improved computing systems as a whole. Indeed, over the years we have observed a steady stream of computing chip designs aimed to increase the processing speed and memory capacity and reduce the physical size. More recently, we have seen a paradigm shift in the input and interaction facilities for devices, moving away from the mouse to support for touch and gesture.

The continuous change and the complexity of the computing ecosystem are important aspects to consider in devising preservation strategies and implementing methods to ensure that the digital artefacts can be sustained and consumed in the future.

Memory Institutions in the Digital Era

Dependence on the ICT Industry

On one side the Information and Communication Technology (ICT) industry is driven by the immediate market needs and focussed on the user consumption of the contemporary computing technologies. On the other, the Memory Institutions, including the Libraries and Archives (L&As), have a mandate to preserve digital assets for future generations. Generally, the preservation process is complex and requires specialists' skills in curation, record management, and handling of content objects. Expanding the mandate from physical artefacts to digital media requires an in-depth understanding of the nature of digital technologies and the ways the Memory Institution can engage with the ICT industry to achieve their objectives.

The past practices of the Memory Institutions reflect the partnerships and collaborations with the main stakeholders and technology providers who are involved in the production and consumption of the printed matter. Currently new alliances are being developed to support the preservation of the digital matter. In order to understand that process, it is instructive to examine the value chain in the production and management of the content in the analogue form and compare it with the production and management of the digital media. It is particularly important to explore the issue of maintaining the integrity of digital artefacts since their nature is radically different from physical artefacts.

Being able to sustain a digital artefact is a starting point of the preservation process. As we will see, handling digital artefacts intrinsically increases the dependence of the L&A practices on the ICT industry. The complexities of the L&A operations, including the record management, risk management, and policy matters are also expected to increase

and should be re-examined. However, without a clear technology strategy to handle digital artefacts and maintain their integrity, it is difficult to make progress on other fronts.

Here we take a high level view on the production, consumption, and preservation of physical verse digital artefacts. The main objective is to reason about the dependencies that the Memory Institutions have on the rest of the ecosystem and how these shape the economic sustainability of the preservation activities.

Value chain in the production and use of physical artefacts

PAPER PRODUCTION →	PRINTING TECHNOLOGY →	CONTENT PRODUCERS →	CONTENT READERS →	LIBRARIES & RCHIVES

Persistence and sustainability of physical artefacts

Printed matter is in the physical form that enables a user with standard capabilities to read and, in normal conditions, store and periodically reuse. Once a book is acquired, there is no further dependence on the book printing technology unless the book is damaged and requires replacement. Through purchase of the book the user secures ownership and consumption in perpetuity. Readers who purchase the book collectively cover the cost of the book production and the book is produced in line with the market demand.

Market forces and demand for content access

L&As assume the responsibility for storing and making the printed matter accessible beyond the point when the demand for purchasing the book diminishes and publishing ceases. In fact the value of the Library service increases when the general demand and supply for book fades away is unfeasible to sustain book printing. The Library is then the only reliable place to access the book. The cost of the borrowing is incurred by the readers and contributes to the sustainability of the library business.

Economic sustainability and ecosystem dependencies

As with any other enterprise, L&As services involve core functions that are provided by the L&As and supplementary functions. For the supplementary functions, they rely upon external service providers. For example, curation of content artefacts is a core activity while the production of shelves, encasing, humidity regulation, and similar technologies are provided by the third parties. Among the latter are technologies that are critical for delivering L&A services but are not core L&A capabilities. For example, modern sensor technologies are used for storage and access to fragile artefacts but the sensor production is beyond the competencies of the L&As. Such dependences on third parties is expected and a source of risk that continuously needs to be monitored and dealt with.

Generally, the cost of the third party services depends on the general market demand and supply and enters the cost structure of L&As services. In instances when L&As are substantial consumers of such services, the L&As become an important factor in the market demand. That creates a more favourable position since L&A's needs will be more readily met while the cost of the production is still spread across the whole market.

At the high level, the digital media production and use involves a similar value chain and similar dependency and risk profiles. However, because of the emerging and fast changing ecosystem, the risk assessments are difficult to make. Furthermore, the nature of the dependencies is much deeper; in fact, intrinsic to the very existence of the content that has to be preserved and delivered.

Value chain in the production and use of digital media

HARDWARE PRODUCTION →	SOFTWARE TOOLS →	CONTENT PRODUCERS →	CONTENT READERS →	LIB. & ARCH.

Computational nature of digital artefacts

In contrast to the physical artefacts such as books, periodicals, and newspapers, viewing digital content requires sophisticated devices and technical infrastructure. Similar to the music recorded on audio tapes, the digital content has to be 'played' every time it is consumed, using specialized devices. In effect, digital documents are 'computed' every time they are instantiated on a screen.

Most of the content productivity software save the output of the authoring process into a file. That file serves as input to the software tools that can process it. Each such tool is, in turn, supported by a stack of other software programs that enable the tool to run on a device. The software stack includes a range of programs, from the operating system to the drivers for input and output devices such as mouse, computer displays, and printers.

In many instances the same file can be processed by several applications and the features of such applications determine how the file is decoded and how output is presented. In essence, the user experience of a digital artefact depends on the software that processes the input file.

Sustainability of digital content

As noted, digital is in essence computational. Instantiating a digital document requires three elements: files with content or data, files that encode the software program, and a computing environment that can run the software program. Storing files can be subjects to errors and experts have developed methods for bit preservation, to minimize the risk of '*bit-rotting*' as content is stored on tapes and disks. However, ensuring that a software program can run presents a challenge. That problem is not new; it is part of the everyday use of software and is dealt with in various ways.

Software providers regularly release upgrades of software programs to ensure that the software applications run properly. This is due to the dependencies and interactions among the programs on the computing devices. Providers continually invest in the software development in order to ensure that their products are usable. That cost is reflected in the price of software and services. While there is sufficient demand for software, it is economically feasible to provide upgrades and fixes. However, once the demands falls, software production is not economically feasible. In that instance, the users are faced with the software obsolescence and unable to access their documents unless such need is satisfied in some other way, e.g., by providers of software that is capable of processing the files. That creates a different digital artefact but may retain sufficient value to the user.

In summary, the persistence of contemporary digital artefact is subject to the economic feasibility of software. The consumers of software directly impact the demand and supply in the software market and their needs and behaviour shape the ICT industry. In the free market economy, the need for long term persistence is met based on the perceived value of the legacy content and the consumers' willingness to pay for such solutions.

The ecosystem dependences

Considering the fundamental dependence of digital content on computation, the task

of preserving digital documents can be defined as *ensuring that some compatible software is available and can run in the contemporary environment to decode and present the content files*. Continuous and inevitable obsolescence of both hardware and software makes this task difficult.

With the mandate to provide long term access to digital documents and to ensure that the content can be experienced by the users, L&As inevitably need to ensure that there exists software that can run in the contemporary environment and process the content. While the printed materials are to a large degree uniform in physical properties and similar methods can be applied to store them, ensure their physical integrity, and provide access to them, the digital artefacts are far more diverse. Thus, the fundamental question is whether there is a similarly 'uniform' and manageable way of having legacy software available in operational form so that they can decode and present legacy files and applications.

'Digital' Bookshelves

Similarly to the physical storage of analogue materials, the 'digital shelves' that hold digital artefacts are provided by specialists using the latest available computing technology. Just as designing and providing physical shelves is not a core L&As business, producing hardware and software components is outside the current L&As competencies. However, the management of artefacts, including storage, curation, search, and delivery are the functions that has shaped the organizational structure and the business models of the L&As.

The basic technical pre-requisites for sustaining digital artefacts include storage of data files, content files, and software files on hard drives, disks, and tapes. Furthermore, it is necessary to provide a computing environment in which the software can run to process the data and content files. That involves acquiring hardware with the appropriate operating system and supporting software. Finally, the preservation practices of both analogue and digital media have been facilitated by digital technologies for creating metadata, searching, and delivering of artefacts on demand.

Since the artefacts are in the digital form, we expect that the characterization of digital objects could be automated to a large extent, in contrast to the characterization of physical artefacts that requires manual entry of information into the content management system. Furthermore, to optimize the internal operations, many L&As have employed their own Information Technology (IT) staff to assist with creating bespoke solutions for characterizing and managing content. Some of these solutions are shared across libraries and archives as common assets. Thus the boundary between the providers of content management solutions and L&As is blurred in this specialist market. By dedicating resources to the development of solutions that are not main stream inevitably changes the cost profile: it provides L&As with more control over the supporting software features and increases the investment into the software development and maintenance to ensure that the software can continue to operate.

Integrity of Digital Artefacts

In contrast to the content management technologies and practices that are applied to both physical and digital media, hosting and accessing digital artefacts are new aspects and practices of Memory Institutions. In order to ensure that digital artefacts can be used it is necessary to store the files of data, content, and software in a manner that all the digital bits are recorded correctly in the storage media. That is a pre-requisite for instantiating the digital artefacts; however bit storing is not sufficient. The preservation systems need to be able to run the software to process the content and data files. The

current practices involve *file format migration, hardware and software emulation*, and *software virtualization*.

File format migration. Migration involves implementation of software that converts files from legacy formats to the formats that can be processed by contemporary software. As the transformed file is used with new software, it is critical to ensure that the new digital artefact meets the objective of the file migration. In many migration practices both the original file and the original software are abandoned, reducing the cost of file storage. However, the migration operation intrinsically *grows at the rate of content production*. In fact, when the newly chosen software becomes obsolete itself, the migration is applied again to previously migrated files and the newly created ones. Finally, file migration is not applicable to digital artefacts that are highly interactive such as simulations and games. An equivalent to the file format migration would be 'converting' the software, i.e., porting the code to run in the new computing environment. In contrast to creating file format converters that are then applied to many content files, code porting is focussed on a single application. Thus, the cost profile is very different.

Emulation and virtualization. Emulation and virtualization aim to provide a computing environment in which digital artefacts can be instantiated in the same way as in the original computing environment where they were created and used in the past.

Emulation, in effect, involves implementing a software program that duplicates the functionality of a computing system, most likely the computing architecture of a computing device or any other hardware component or software that may be required to instantiate and use a given digital artefact. Being a software itself, emulator is developed for a specific operation system (OS) and thus needs to be updated as the OS becomes obsolete.

Virtualisation, on the other hand, is a software layer above the physical hardware that allows guest Virtual Machines (VMs) to access computing resources of the host machine. VMs are built on top of the abstraction layer, called *hypervisor* which acts as a traffic controller and coordinates the access and use of the physical resources. The VM technologies evolve with the computing architectures of the host machines. They are primarily built to support specific operating systems, or to ensure sufficient support for running individual applications. Thus, they evolve with a demand for virtualizing a particular generation of software. Furthermore, the VM development may be guided by the requirements for running legacy systems in a safe and secure manner, making provisions for security breaches if the 'unsecure' legacy applications need to connect to external resources, such as the Internet or the Cloud.

The cost of emulation and virtualization is primarily related to the development and maintenance of emulation and virtualization software due to the changes in the computing ecosystem. While the core cost of emulation and virtualization does not necessarily decreases, it is independent from the rate of content production. The number of VMs that need to be provided depends on *demand to access digital content* which could be exploited for revenue generation.

Digital Preservation and the ICT Ecosystem

Industry Trends and Digital Preservation

The key to the economic sustainability of any initiative or enterprise is its relationship to the rest of the ecosystem. While the ecosystem can provide what the enterprise requires and while the enterprise generates value that can be exchanged within the ecosystem, the enterprise will be sustained.

Any digital content owner has a fundamental dependence on the computing industry. For the Memory Institutions that is reflected in the basic need for (1) hardware and software to store data, content, and program files, (2) hardware and software to run applications

and instantiate digital artefacts, and (3) hardware and software to run the preservation processes, from preservation planning to ingest, metadata generation and customer facing services. This implies a need for risk assessment in a range of specialty areas of computing, from tracking developments in the computer architectures and storage devices to the trends in the design of content management applications and supporting information architectures.

When the global market is not responsive to the specific needs of an organization, it is often necessary to invest in bespoke solutions. Unless the organization has staff with the software development competencies, that task is typically outsourced to a third party. In that instance, the dependence is not necessarily on the commodity software providers but on the providers of services and skilled labour. The cost of bespoke systems is expected to be higher than the cost of technologies in the main stream market. However, if the components of such systems are produced as part of a broader need, the integrated system can be built at a reasonable cost. Two current trends in ICT industry are particularly favourable for making digital preservation activities integrated and thus supported by the ecosystem. These are the shift to the *cloud computing paradigm* and the emergence of *Big Data* management analytics.

Cloud Platforms and Preservation Services

Cloud platforms comprise large computing centres that are managed and optimized for data storage and computation. They are conceived to maximize the effectiveness of the shared resources. The resources are dynamically reallocated based on demand and that increases the user's ability to re-provision resources as needed. Through automated metering of the cloud usage, the user can build a model of utilization and plan provisioning based on the cost and use. Furthermore, with the use of multiple redundant sites, one can increase reliability, ensure continuity, and support various strategies for disaster recovery.

Fundamental to the cloud computing is virtualization. Through virtualization, one can 'partition' a physical computing device into one or more virtual devices, each used and managed to perform specific computing tasks. Through virtualization at the operating system level, one can create a scalable system across multiple independent computing devices.

Considering the digital preservation requirements, there is a perfect alignment between the cloud approach, based on virtualization, and the need to host and run legacy software in order to instantiate digital artefacts. Therefore, technologically, the preservation needs can be met by the main stream cloud platforms. It remains to create services that are economically sustainable. Fortunately, cloud platforms are offered in different forms, providing a range of choices.

Cloud Services Models and Architectures

The cloud providers offer their services based on several models:

Infrastructure as a service (IAS)—the most basic cloud service that offers physical or virtual machines to the user. The cloud users typically install operating system images and applications on the cloud infrastructure and maintain them. The service may offer additional resources such as firewalls, load balancers, virtual local networks, and similar.

Platform as a service (PAS)—the service that provides computing platforms to the user, including the OS, support for running programming languages, web servers, and databases. This is suitable for application developers who can develop and run software on the cloud without buying and managing the underlying hardware and software layers.

Software as a service (SAS)—the service provides access to applications and databases on demand. Thus, the cloud provider manages both the infrastructure and the platforms that run the software, and typical apply the pay-per-use model.

With regards to the management of cloud resources, there are different infrastructure models. They primarily vary based on the control and exclusivity that an organization or a user may want with regards to the infrastructure, the services and the communication protocols between the devices and the cloud. Typically, there is a trade-off between the level of control and the investment required to uphold that control by the organization, considering the ever changing computing ecosystem. The control often includes the concerns related to the trustworthiness and stability of the infrastructure as well as the security, data protection, and privacy of the access and communication protocols. Among common models of the cloud infrastructures are:

Public Cloud—the cloud service is provided over a network that is open for public use. For example, Amazon AWS, Microsoft, and Google offer the computing infrastructure at their data centres and, by default, enable access via the Internet. Private connections can be purchased or leased using, for example, "AWS Direct Connect" and "Azure ExpressRoute". Management and maintenance of the infrastructure is the responsibility of the cloud providers who provide cloud services using different models, e.g., IAS, PAS, and SAS. as discussed above. Associated with them are different cost and payment models. It is typical to use cloud services through a subscription and a pay-as-you-go. That assumes a moderate commitment to the cloud services and a provisional plan where to put the data in case the organization ceases to use of the cloud infrastructure.

Private Cloud—the cloud infrastructure is created solely for the use of a single organization. It can be managed by the organization or outsourced to a third party; it can be hosted within the organization or externally. In this instance the business takes on the responsibility to virtualize the business environment in the cloud, keep abreast of the technology developments and re-evaluate decisions about IT resources. Because of the need to manage them closely by the organization, they do not provide the benefits of the computing abstractions that public clouds have. Technically, there may be little difference between the private and public cloud except for the security requirements.

Community cloud—the cloud infrastructure is shared between several organizations with common interests and similar requirements, such as privacy, security, and compliance. Similar to the private clouds, they can be hosted and managed internally or by a third-party. In this instance, the cost is spread among multiple organizations and thus the savings are larger than in the private cloud, although lesser than the public cloud.

Hybrid Cloud—the cloud service is a composition of models, such as private, public and community clouds that remain distinct but bound together. They offer benefits of multiple deployment models. The hybrid cloud services aim to bridge offerings of different providers and allow the user to extend the capacities and capabilities by integrating, aggregating, and customizing constituent cloud services.

Cloud Services and Digital Preservation

In the context of digital preservation, the cloud services are often debated because, at this early stage of the cloud computing industry, it is hard to assess the risk models associated with the use of the cloud services. The issues involve pricing, quality of service, security and privacy, and sustainability over long time.

With regards to the latter, cloud services are subject to the global market forces and the same economic issues that make the computing systems obsolete apply to them. However, there is historical precedence of services that become fundamental to the society, such as water, gas, and electric grids, and therefore less vulnerable to the economic factors. Therefore, if the cloud services follow the same commoditization path, the long term technical sustainability may not be an issue.

All other aspects are less specific to the digital preservation agenda and common to other industries. The price, in particular, will vary for different models of cloud computing and be subject to the market demand. The discussed models of the service provision and the infrastructure management provide a range of possibilities for the Memory Institutions to choose from. Fundamentally, each institution will have to decide on the level of control and responsibility for the computing infrastructure and, based on that, plan the budgets and internal processes.

Cloud Computing and Big Data Analytics

In parallel with the shift to the cloud computing, we observe a strong emphasis on Big Data analytics where large and complex data sets are processed using hundreds or even thousands of servers. In many instances the analysis has to be done over historical data and therefore information needs to be extracted from different file formats. That, in turn, requires file characterization and content extraction, very much in line with the processes included in the preservation workflows. Therefore, the main stream Big Data analytics industry can benefit from the tools developed for preservation and vice versa.

More importantly, the Big Data analytics is increasing the demand for infrastructures, tools, and services to support management and processing of large repositories of heterogeneous digital assets. Consequently, the cost and quality of file storage and software virtualization are being optimized, directly benefiting the preservation initiatives. Moreover, the practices and preservation are likely to become an integral part of the economically viable 'Big Data' businesses.

Digital Heritage and the Role of UNESCO

Voice of the society. Preserving legacy data for the posterity has intangible and deferred value. Therefore the sustainability models are typically driven by the notion of the *common public good* and supported by the society as a whole. In instances when preserving legacy becomes tightly integrated with commercial operations, the cost can be absorbed directly by the enterprises that generate economic gain. In the former case, the UNESCO has played a critical role in representing the voice of the societies and raising the need for action at the international level. In the case of digital preservation, UNESCO continued to be a unifying force for Governments and Memory Institutions to scope the problem and pave the way to the effective solutions.

Facilitator. In this paper we have discussed the technical feasibility of digital preservation and the technology developments that are in favour of creating effective solutions. While such considerations are fundamental to the preservation efforts, they have to be complemented with concerted efforts to:

- Facilitate partnership with the computing industry and secure rights to use computing technology beyond the point of economic viability within the original business model. This includes the licensing of legacy software and the management of IP rights related to the design of hardware, services, and applications.

- Promote research programmes to investigate engineering practices that take into account the full life-cycle of computing technologies, from the research and commercialization phase to the sustainability for common good. In the latter stage it is particularly important to consider how to make the maintenance and repairs technically feasible.

- Support the specification of standards that follow the evolution of the computing innovation and the requirements that arise from the new digital technologies.

For example, recent developments towards dynamic and highly interactive digital artefacts challenge the assumptions of traditional preservation approaches that focus on files. That forces us to expand our focus from file storage and bit preservation to creating

environments with running applications. Thus, it is necessary to define a standard for the *preserved digital record* as a hosted virtual component that encapulates the software required to instantiate the digital artefact and to access its metadata.

Custodian. The computing ecosystem is complex and fast moving. Thus, in order to ensure that the digital preservation can be undertaken by the Memory Institutions and that supporting service can be established, it is important to put in place facilitating functions. We have already mentioned the need to establish partnership with the computing industry. It would be equally important to create a number of facilities that enable implementation of the preservation solutions, including:

- The bank of hosted legacy software that is running as part of the contemporary platforms. That will serve as validation that the legacy content is accessible
- The documentation of legacy software and file formats
- The bank of legacy format transformers and, when necessary, software source code to enable the maintenance and repairs of the hardware and software systems.

Considering the favourable technology developments in the computing industry and the momentum that UNESCO has already created, we are on the way to address the digital preservation in a holistic and sustainable way.

3. Beyond a software and hardware repository

by Michiel Leenaars
Stichting NLnet

Summary

In this short contribution the author posits that due to the proliferation and diversification of use of software and hardware solutions involved with documents, documents typically will have mixed origins. The often destructive handling of non-default content elements leads to unpredictable outcomes, which is a structural problem that cannot be solved without improving the software. This disaster can be compared to the acidic inks that threaten cultural heritage of a significant part of the 20th century, and can only be solved by improving the characteristics of the applications. The repository approach - storing legacy software and hardware in the hope they can be used for future access to specific documents - should be repurposed to find problematic issues through automated interoperability testing.

Introduction

Long term access and trustworthy preservation of digital information cannot be realistically achieved if already at the time of creation the path to access the content involved is as narrow as a single specific version of an application tied to a specific combination of hardware and software platforms. Historically, the problematic nature of this issue is under-appreciated. In the modern internet era things are even worse; what seems an application may consist of many. Parts of the functionality of an authoring, editing or reading tool may be tied to a (cascade of) remote on-line services that all need to be available and accessible in conjunction. Such a scenario for instance may happen when a user is interacting with remote objects such as fonts or scripts, or in the case of parts of the software or parts of a document involving certain so called 'digital rights management' solutions. Unlike its paper equivalent, digital rot may be completely invisible.

Of course there is the possibility to try and capture both the old and the new reality by creating an exhaustive repository of snapshots and versions of software and hardware from a multitude of vendors. That way, at any point in the future one could fire all these up in software containers and exhaustively try each and every combination when needing to access specific historical content of interest. While certainly there is not much against having such a repository as a fallback in case all else fails, it doesn't address the most serious issues. In fact, there are already a number of assumptions that threaten reliability even in the here and now. It is not necessary to wait for these problems to emerge, as one can already witness them in present day common usage.

The first issue is that (for someone other than the original author) there is really no way to know or validate whether or not the right match is made and the full content is shown. Any assumption about another technical setup, even with close version of the software available, may be wrong - because the programmers are actually touching the sensitive parts of applications that deal with precisely the parsing of content and any difference may cause a butterfly effect. Even a single commit from a programmer may invoke a regression. And because applications are optimised to use advanced hardware features in CPU architectures, such as in the case of spreadsheets that can profit a lot from hardware assisted computations, applications may even be affected by the microcode version of a specific CPU used.

Most users have learned the hard way that applications tend to be incorrect and incomplete in reporting their actual ability to handle all the elements present in a document with care. There are no systematic tests available that actually prove that one

application is able to access all the content produced by some other (version of the same) application. And unlike the situation of a paper archive, where everything is in plain sight, digital objects can go missing - or hide in plain sight without expert understanding of the application to access specific features. The author has witnessed a number of instances where features required simultaneous, non-intuitive handling of two different input devices (e.g. track pad and keyboard) to invoke the intended action.

In an ideal world, the documents users send out and receive are all instances of a non-ambiguous, perfect document model. The model is proven complete, described in full immutable mathematical glory as if the specifications of the underlying markup were chiseled in stone for eternity. Operations are fully transparent and commutative, hence a well-behaving application could not possibly lose any information the user entrusts to it - and the interface is entirely consistent in how it maps to user intentions. Applications even have no way to alter, modify or skip any existing elements they do not understand - because there is no such thing. Software could completely predictable in moving from one document state to the next. The final markup produced by any series of actions in the user interface is identical to what anyone armed with a file format specification and the same set of operations would create on paper, and vice versa. The document model is identical to the internal representations of the document model.

Of course, that is not what the reality of documents and their life-cycle looks like. In the actual world software was created by humans rushing their products out to meet deadlines, based on incomplete or flawed understanding of the problem space. Developers live from patch to patch, from release to release - each seeing only a part of the entire application. The requirements from supplier to supplier vary wildly - from small screen mobile usage with limited CPU and touch interface, web applications that have to deal with the quirks of many different browsers and peculiarities of for instance Javascript, special needs such as voice and gesture interaction and much more. Sometimes developers depend on platform-specific libraries, which means they write custom glue code to integrate these and inherit different behaviour on different platforms. In order to 'make things work' under these vastly different circumstances, developers often have to include crude 'temporary' hacks to make things work slightly better than not, based on incomplete reverse engineering of the work of others. They asked to consciously copy some known historical errors to not break existing documents, and are unaware of other errors.

Add to this chaotic interaction the diversity in usage inherent to complex applications. If you sit next to another person editing a longer document beyond the most basic operations, you will at some point likely experience some friction by their idiosyncratic interaction with the user interface. Even within a single application using the exact same hardware, different users may create vastly different documents - using a personal blend of direct formatting in combination with styles, and using their own personalised workarounds, macro's and other technical shortcuts. Since no document is created alike, complex effects may occur when different applications or slightly different versions of the same application are used together and create interference at the markup level unique to an individual.

We really don't know how bad this problem is, as there is no decent infrastructure to share knowledge or test for specific behaviour. We do know there is no going back: once documents start appearing in the wild, everybody has to deal with each other’s hacks, approximations and restrictions. As such storing standalone tools is but a partial answer, because different parts of a document are actually likely to come from different origins during its lifetime. And to make it worse: during that path, each host application has been free to modify, transform or delete any underlying element. No matter how subtle these differences may be, one can count on any application to overwrite the authoring tool metadata and thereby obfuscate a documents' past. Current office applications are technically incapable of keeping a complete history of changes at the markup level, although elegant technical solutions to do so have been proposed. The suppliers tell the

standards committee that software simply isn't capable enough: legacy application can't even keep the name of the previous-but-one editing tool used, let alone track changes at the level of a single character. Yet that can be all it takes to evoke different behaviour from a consuming application.

One doesn't have to dig too deep for problems to start emerging. The various document interoperability events with vendors and the community organised by OpenDoc Society (the so called 'ODF Plugfests'), have so far made it quite clear that many issues exist and it is worthwhile to highlight them. Depending on the content, merely opening and immediately saving a document in an office application without as much as even touching a button can and will already significantly 'damage' a document in quite a few cases, and with every modification of the new document the distance to the previous content increases.

Cross-application usage, long term access and technical consistency did not sit at the heart of the design process of the applications we use. And certainly interoperability is a moving target that depends on a healthy group equilibrium rather than on individual endeavours. Reaching such an equilibrium depends on cooperation between all stakeholders, and yet under the current circumstances there may be perverse impulses for major actors to promise interoperability but secretly steer away from it. Lack of interoperability can have significant commercial value, as uncertainty about the ability to reliably access ones historical content makes it difficult for users to switch away from a specific application. Indeed, users are known to stick to suppliers for as much as decades despite huge dissatisfaction with prices. In other words: incompatibility discourages the use of competing products, and benefits the currently dominant players. There have been cases of major companies in the office application area being indicted in court for putting software on the market that emulated fake errors when noticing the involvement of other vendors. With billions of annual profit involved, the stakes are high.

Whether or not applications maim documents intentionally, the user gets the short end of the stick. The gravity of the problem can be compared in scale to the issue with inks used in book printing in the middle of the 20th century, where acids present in the ink would eat away the paper - with the difference that with electronic media the effect is less visible and could already start happen immediately when a document is handed over to another device or user. It is a fact that office applications at present have no problem to silently shred document elements and metadata they do not understand, whether these were created or added by another application or by a different, incompatible version of the same tool (or its dependencies). As an example: it appears some office applications are not capable of handling frames within a frame, and will irreversibly convert any such objects into images upon merely opening a document - overwriting the original content. Visually this may look completely indistinguishable, but this loss of information in the handover between authors has severe consequences later on because visually impaired users will no longer see the contents and text search no longer works. Another example: some office applications are incapable of handling RDF metadata as specified in the OpenDocument Format 1.2 specification, and will typically lose all the (valid) metadata contained in a document without giving any feedback to the user.

The need to solve the challenges of long term content access puts the heritage and archiving world into the no man’s land that not too long ago was called the 'office format war'.

The way forward

Working towards a true long term solution for accessible content starts with the acknowledgement that modern documents are indeed hybrid creations, the unique product of multiple tools used at different stages by different users in different environments. For archiving it therefore is just not good enough to conserve the 'original application' in which documents are created. What does such a label mean if a document

initially produced in Wordperfect, later edited and saved in the legacy binary file format of Microsoft (still within Wordperfect), and subsequently mailed back and forth between a user of AbiWord on a Linux laptop and another author using LibreOffice on Mac OS X utilising the OpenDocument Format? We can't expect the official file formats to adequately describe such documents, given the abundance of unpredictable combinations at the markup level that could lead to interference as well as the individual mappings applications have between different document models.

In order to improve future interoperability and safeguard future access, actual convergence is needed as well as insight in the actual individual quirks of applications. Both can only happen based on actual shared empirical knowledge how applications deal with each other's output as input. First of all the relevant applications need to be mutually available for testing to developers. The repository that the PERSIST project is aiming for would be an excellent facility for that purpose. During the ODF Plugfests mentioned earlier, OpenDoc Society has already successfully performed experiments with an internet based tool called Officeshots. Officeshots offered online access to multiple versions of multiple office applications on multiple platforms in parallel, coupled with multiple automated validators and other tests that would be run upon submission of a document. Users could submit their own custom documents, to test what would happen if these would be opened in the various applications. This worked in batch mode, so that a single upload was enough to see if the output led to conformance issues in all applications present. (Sets of) documents could be made persistent in a public gallery, so they would automatically be tested against future versions once these would become available. Vendors could add their own development version of an application to the set of applications on the fly while it was still being written, to allow others to test against it with their documents without having to update many instances or exposing the actually unreleased application to the outside. Especially with major feature releases such early access is attractive for early testing for upcoming incompatibilities - once an application is shipping to customers, all that is left is damage control.

All that is required to perform structured testing is a set of applications available online and a unified API to interact with them. Officeshots as a first experiment offered roundtripping documents (open a document and immediately save it again), exporting screenshots and generating (digitally) printed versions of a document through the connected applications. The volunteers that ran Officeshots over time provided nearly a hundred different variants of Office applications, using native facilities of some applications (some of which were developed especially for the purpose of facilitating Officeshots), using custom software runtimes (such as the open source OfficeConvert tool that also was created especially for this purpose to handle the automation of Microsoft Office), by scripting the application with native scripting abilities or with GUI scripts that used the accessibility layer of the operating system to script document manipulations.

The Officeshots experiment proved very useful for both debugging purposes for developers as well as understanding for users. For the purpose of long term accessibility of information the ideas from the Officeshots experiment could be taken further. Officeshots as a beta level project already took significant steps towards addressing the problem by including the ability to have persistent test suites, but it lacked the scale to fulfill all the possibilities that approach had.

The idea is simple: documents contain a limited set of nestable elements, possibly with layers, foreign objects, fallbacks, metadata, annotations and signatures. By using this knowledge and more specifically the technical specifications of the various document formats, one could create a feature complete synthetic test suite for each relevant file format as a starting point for research. A synthetic document test suite can exhaustively provide atomic tests with each possible combination of elements, up to a predefined depth of recursive nesting of course. From the most basic documents with a paragraph of text or a single cell, all the way up to an unrealistic but valid document with an 8 bit greyscale TIFF image inside an image frame that is anchored within a heading within a

footnote on a page with a large table containing a text frame with an underlined quotation with tracked changes and RDF metadata inside another table docked in an image frame anchored to a page with sections - etcetera. The practical parameters for how deep one should go with this stress test, could be probably be qualitatively derived from a representative set of real world documents. Testing for anomalies in handling the elements involved can easily be automated exactly because the documents are still so simple that they can be generated algorithmically. Up till which point are objects still intact after a document has been roundtripped by the various applications, and if not - by comparing with other documents investigate what exactly happens. The roundtripped documents themselves become part of a new stage of testing - because subtle modifications of an application during the import-export cycle which still result in valid markup, may produce unforeseen consequences later on in other applications. As the amount of possible combination goes up fast with every following stage, there may be a practical limit.

Research as described above will allow to create 'behavioural fingerprints' of legacy applications: elements that show typical mutations (however minute) resulting from contact with a certain (version of an) application. Sometimes the actual element may have already disappeared because of such an encounter with fate, but perhaps one can identify remnants of the context elsewhere in the file which may indicate the historical presence of it nonetheless.

A synthetic test suite should be complemented with a crowd sourced test set of documents typical for certain work flows, or already known to cause interoperability problems. Officeshots provided a 'greeking service' which allowed the community to easily submit internal documents that were not suitable for full publication but had interesting characteristics (for instance resulting in unknown errors in specific applications).

With the help of the test results from running the test documents through all applications, real world documents can be turned inside out - subtract known constructs that can be expected to work and what you have left are unpredictable constructs as they appear in actual documents. These can be isolated for research and also fingerprinted as well, and then fed to the online repository handlers.

This will allow for iterative qualitative mapping of any known problematic construct inside what was previously a dumb bit bucket. An archive of fingerprints and indicators will potentially allow to automatically perform assistive digital archeology on any incoming document in an archive. Documents can be 'sniffed' to see if there is reason to suspect potential loss, misplacement or damage. Features that are less common can be pointed out to the researcher automatically. The document archive may even attempt to perform a reconstruction of the documents' history to inform the researcher of possibly successful avenues ('We suggest using both TextMaker 2012 on Linux and Calligra Words 2.8.5 on Windows to view this file') and inform the researcher of possible risk of loss or modification that may have occurred along the journey from cradle to e-depot. Researchers can use the information to seek similar documents or documents with identical content elements inside large collections of documents, to investigate how certain documents came into being. In the data explosion researchers have to live with, such assistance is invaluable.

Coda

For long term preservation the industry really needs to clean up its act, likely with the help of tooling similar to what is described above. The act of producing documents that can only be accessed in a single application, is not sustainable. While we can mitigate some of the issues, there is a significant social and financial cost attached to that which can become inhibitive. UNESCO as a major stakeholder can help to make the actual problem visible, as well as making the vendors directly accountable for putting the content of their users at risk - both in the here and now, as well as in the future.