Digital Preservation **Handbook**

Technical Solutions and Tools

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Who is it for?

Operational managers (DigCurV Manager Lens) and staff (DigCurV Practitioner Lens) in repositories, publishers and other data creators, third party service providers.

Assumed level of knowledge

Novice to Intermediate.

Purpose

- To focus on technical tools and applications that support digital preservation: software, applications, programs and technical services.
- To consider the practical deployment of preservation techniques and technologies whether as relatively small and discrete programs (like DROID) or enterprise wide solutions that integrate many tools.
- This section excludes other more strategic or policy issues and standards that are sometimes described as tools: these are covered elsewhere in the Handbook.

Gold sponsor The National Archives

Silver sponsors British Library Jisc ARA Archives & Records Association UK & Ireland dpc

Bronze sponsors National Records of Scotland Charles Beagrie

Reusing this information

Please use this form of citation for the Handbook: Digital Preservation Handbook, 2nd Edition, http://handbook.dpconline.org/, Digital Preservation Coalition © 2015.

Contents

Tools

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

A beginner's guide to digital preservation tools

The utility of technical tools for digital preservation depends on the context of their deployment. A community recommendation may be strong but if it does not align with your specific function or organisational context then there is a significant chance that the tool will fail to perform. So before selecting digital preservation tools it is important to consider carefully the technical workflow and institutional setting in which they are embedded. A practical example of this has been presented by Northumberland Estates who developed a straightforward evaluation framework to assess tools in context.

An alternative way to consider this topic is to review the extent to which any given tool will deliver preservation actions arising from an agreed preservation plan, which in turn derives from a given policy framework.

Thinking about digital preservation tools

The following issues are frequently encountered in the process of deploying digital preservation tools. This is not a comprehensive list but consideration of these issues will help sensible and realistic choices.

Open source versus commercial software

Some organizations - often in higher education and especially institutional research repositories - are comfortable with the use of open source software, especially where they have an in-house group of developers. 'Open source' software is where the underlying code is made available for free, enabling a free flow of additions, amendments or development. Other organizations which don't have easy access to developers, tend to have procurement rules that prefer 'off-the shelf' commercial solutions backed by on-going support contracts. The distinction between Open Source versus Commercial software is often over-stated because both influence each other. Nonetheless you may need to consider your organization's norms and culture while you select tools.

Enterprise-level solutions versus micro-services

Some digital preservation tools are designed to offer 'soup to nuts' solutions, meaning that they provide an integrated end-to-end process that enables all (or most) digital preservation functions to be delivered for a whole organisation. In fact enterprise-level solutions are most often constructed by aggregating individual tools integrated into a single interface. The solution to any given problem might be relatively simple and your organisation may be happy assembling a series of small tools for discrete functions. This encourages rapid progress and is helpful with testing and trialling tools; but it can be hard to maintain over an extended period. In other organisations there is much tighter control over the deployment of software and an expectation that solutions are built across an entire workflow - requiring comprehensive solutions. This can be slower to respond but can be more sustainable in the long term. Before selecting a tool it is helpful to consider where on this spectrum your organization normally sits.

Describing workflows

A key consideration for tools is where they sit on an overall workflow so before selecting tools it helps to consider and map out the entire workflow. Being explicit about a workflow can also help identify redundant processes as well major bottlenecks. One frequent challenge is that tools solve a problem in one element of a workflow, only to create a problem elsewhere. In addition, organisations may have multiple workflows that may have different requirements that conflict in some way. Describing a workflow therefore provides a basis for anticipating difficulties and can provide a roadmap for ongoing development.

Specifying clear requirements

In order to evaluate the usefulness and value to your organisation of the many tools available it helps to have an explicit statement of requirements. Tools can be compared and benchmarked transparently and decisions justified accordingly. Properly executed, requirements-gathering activities can involve a range of stakeholders and therefore maximise the potential for alignment and efficiency, achieving wider strategic and organisational objectives.

Changing and evolving requirements

It is normal for requirements to change through time. Indeed digital preservation is largely concerned with meeting the challenges associated with inevitable changes in technology. So it is necessary to monitor and review tools to ensure that they remain fit for purpose and that any changes in requirements are made explicit. A periodic review of the specification of requirements is recommended.

Sustainability of tools and community participation

An important consideration in any decision over the tools you use for digital preservation is the sustainability element. Sustainability in terms of tools may include an active user base, support, and development. For instance, a large user base, both in terms of commercial and open source providers can be a vital indicator for identifying a viable tool.It's worth noting that a community can change rapidly and for reasons that might not be easily predicted. 'New kids on the block' can quickly become mainstream while large communities can dwindle as quickly as new technologies overtake existing ones. Consequently it may be necessary to monitor the health of the developer community supporting your tools.

Finding digital preservation tools: tools and tools registries

One of the welcome features of digital preservation in the last two decades has been the rapid development of software, tools and services that enhance and enable digital preservation workflows. As the digital preservation community has grown in size and sophistication so our tools have become more powerful and more refined. This proliferation and increased specialism can also act as a barrier to deployment: especially when tools have been the product of relatively short lived research projects with limited reach. Consequently the diversity of tools can seem increasingly bewildering to new users, while the route to market for developers is increasingly complicated.

Tools registries have emerged in recent years as a way to help users find tools that they need. A number of registries now exist that describe digital preservation tools. Depending on the interests of the people behind them, they can also provide detailed descriptions, reviews or comments about tools from the wider community. So they are not just helpful for users: by allowing experts to review tools and assess their performance they signpost strengths and weaknesses and provide a basis for future development; by connecting tools to users they help developers reach a much wider audience and get feedback to improve their tools.

Registries are a common way for the digital preservation community to share information. Other types of registries exist such as 'format registries' that outline the performance of given file formats, or 'environment registries' that describe the technology stack necessary to create an execution environment to emulate or virtualize software. These are covered elsewhere in the Handbook.

Too many registries?

While registries are a good way to manage the proliferation of tools, it is now recognised that a proliferation of registries is also a potential barrier to use. The COPTR registry was designed specifically to address this problem, drawing on data from multiple sources including DCC, POWRR, and the Library of Congress.

Practical support and guidance

Having considered some of the tools registries and digital preservation tools that are available to organisations, the next question that often arises is which one to choose that fits your organisational purpose. First and foremost it is important that your selection is aligned to organisational need and strategic direction; the resources and case studies below provide evaluation tools and advice to support successful implementation.

Resources

Tool registries

Community Owned digital Preservation Tool Registry COPTR

http://coptr.digipres.org/Main_Page

COPTR describes tools useful for long term digital preservation and acts primarily as a finding and evaluation tool to help practitioners find the tools they need to preserve digital data. COPTR aims to collate the knowledge of the digital preservation community on preservation tools in one place. It was initially populated with data from registries run by the COPTR partner organisations, including those maintained by the Digital Curation Centre, the Digital Curation Exchange, National Digital Stewardship Alliance, the Open Preservation Foundation, Preserving digital Objects With Restricted Resources project (POWRR) http://digitalpowrr.niu.edu/ listed below. COPTR captures basic, factual details about a tool, what it does, how to find more information (relevant URLs) and references to user experiences with the tool. The scope is a broad interpretation of the term "digital preservation". In other words, if a tool is useful in performing a digital preservation function such as those described in the OAIS model or the DCC lifecycle model, then it's within scope of this registry.

APARSEN tools registry

http://www.alliancepermanentaccess.org/index.php/tools/tools-for-preservation/

The APARSEN tools repository attempts to build an evidence-base for preservation tools, and in particular to try to identify which tools are appropriate for which type of data. APARSEN collects details of preservation related software, examples of data, and the evidence of preservation linking software to types of data. Some of this evidence comes from specific testbeds but much comes from user scenarios. The resource is now maintained by the Alliance for Permanent Access (APA).

AV Preserve tools list

http://www.avpreserve.com/avpsresources/tools/

A list of tools of particular use in the long term preservation of audio visual materials, both digitised and born-digital.

Digital Curation Centre (DCC) tools and services list

http://www.dcc.ac.uk/resources/external/tools-services

The DCC is a centre of excellence, to support researchers in the UK tackling challenges for the preservation and curation of digital resources. To achieve this goal it offered a number of support and advisory services supported with targeted research and development. The former includes a catalogue of tools and services which categorises tools for researchers and curators. The information is also integrated in COPTR (see above).

DCH-RP registry

http://www.dch-rp.eu/index.php?en/137/registry-of-services-tools

The Digital Cultural Heritage Roadmap for Preservation (DCH-RP) tools registry collected and described information and knowledge related to tools, technologies and systems that can be applied for the purposes of digital cultural heritage preservation. Version 3 of the registry was created in 2014.

Inventory of FLOSS (Free/libre open-source software) in the cultural heritage domain

https://docs.google.com/spreadsheet/ccc?key=0Ag_7rVJwt0CpdFRJOEJxdEk4ZEMxQ01jaDgxQXFSTkE#gid=0

Produced by the EU funded Europeana Project, this inventory lists free open source software which may be of use in the cultural heritage sector. While not limited to digital preservation tools the inventory does contain information on a variety of tools with digital preservation applications, assessing their purpose, quality of documentation, level of support, license requirements and providing links to project information and source code. Background information on FLOSS is available on the Europeana site http://www.europeana.eu/portal/.

Library of Congress NDIIPP tools showcase

http://www.digitalpreservation.gov/tools/

The Library of Congress's digital preservation tools registry is a selective list of tools and services of interest to those working in digital preservation. It is no longer being actively maintained and content is integrated in COPTR (see above).

Preserving digital Objects With Restricted Resources (POWRR) Tool Grid

http://digitalpowrr.niu.edu/tool-grid/

POWRR investigated, evaluated, and recommended scalable, sustainable digital preservation solutions for organisations with relatively small amounts of data and/or fewer resources. A significant output of the project was the tool grid produced in early 2013 based on the OAIS Reference Model functional categories. An up to date version of the POWRR Tool Grid can now be generated in COPTR (see above).

Digital Preservation Q&A

http://qanda.digipres.org/

This is a site where you can post queries and answers to help each other make best use of tools, techniques, processes, workflows, practices and approaches to insuring long term access to digital information. Digital Preservation Q&A is currently moderated by representatives from NDSA and OPF member organizations.

Practical e-records

http://e-records.chrisprom.com/author/prom/

Software and Tools for Archivists blog from Chris Prom. Although some information may be several years old the blog provides a useful starting point for understanding the uses of a variety of tools for digital preservation and a standardised evaluation of the tools against set criteria, including ease of installation, usability, scalability etc. In addition to information on tools the blog contains a host of other useful resources, including policy and workflow templates, recommended approaches.

Case studies

Diary of a repository preservation project

http://blog.soton.ac.uk/keepit/

A record of progress (between April 2009 and September 2010) as the Jisc-funded KeepIt project tackled the challenges of preserving digital repository content in research, teaching, science and the arts. It includes helpful experience for assessing preservation tools.

Northumberland Estates

http://wiki.dpconline.org/index.php?title=Northumberland_estates_case_study

Northumberland Estates developed a straightforward evaluation framework to assess tools in context. The project set out to survey digital repository options currently available for small to medium organisations with limited resources. Note the recommendations reached in the final business case reflect the organisational needs of Northumberland Estates and may not align themselves with your own goals. The case study was prepared as part of the Jisc-funded SPRUCE project.

Fixity and checksums

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Fixity

"Fixity, in the preservation sense, means the assurance that a digital file has remained unchanged, i.e. fixed." (Bailey, 2014). Fixity doesn't just apply to files, but to any digital object that has a series of bits inside it where that 'bitstream' needs to be kept intact with the knowledge that it hasn't changed. Fixity could be applied to images or video inside an audiovisual object, to individual files within a zip, to metadata inside an XML structure, to records in a database, or to objects in an object store. However, files are currently the most common way of storing digital materials and fixity of files can established and monitored through the use of checksums.

Checksums

A checksum on a file is a 'digital fingerprint' whereby even the smallest change to the file will cause the checksum to change completely. Checksums are typically created using cryptographic techniques and can be generated using a range of readily available and open source tools. It is important to note that whilst checksums can be used to detect if the contents of a file have changed, they do not tell you where in the file that the change has occurred.

Checksums have three main uses:

1. To know that a file has been correctly received from a content owner or source and then transferred successfully to preservation storage
2. To know that file fixity has been maintained when that file is being stored.
3. To be given to users of the file in the future so they know that the file has been correctly retrieved from storage and delivered to them.

This allows a ‘chain of custody’ to be established between those who produce or supply the digital materials, those responsible for its ongoing storage, and those who need to use the digital material that has been stored. In the OAIS reference model (ISO, 2012) these are the producers, the OAIS itself is the repository, and the consumers.

Application in digital preservation

If an organisation has multiple copies of their files, for example as recommended in the Storage section, then checksums can be used to monitor the fixity of each copy of a file and if one of the copies has changed then one of the other copies can be used to create a known good replacement. The approach is to compute a new checksum for each copy of a file on a regular basis and compare this with the reference value that is known to be correct. If a deviation is found then the file is known to have been corrupted in some way and will need replacing with a new good copy. This process is known as ‘data scrubbing’.

Checksums are ideal for detecting if unwanted changes to digital materials have taken place. However, sometimes the digital materials will be changed deliberately, for example if a file format is migrated. This causes the checksum to change. This requires new checksums to be established after the migration which become the way of checking data integrity of the new file going forward.

Files should be checked against their checksums on a regular basis. How often to perform checks depends on many factors including the type of storage, how well it is maintained, and how often it is being used. As a general guideline, checking data tapes might be done annually and checking hard drive based systems might be done every six months. More frequent checks allow problems to be detected and fixed sooner, but at the expense of more load on the storage system and more processing resources.

Checksums can be stored in a variety of ways, for example within a PREMIS record, in a database, or within a ‘manifest’ that accompanies the files in a storage system.

Tool support is good for checksum generation and use. As they are relatively simple functions, checksums are integrated into many other digital preservation tools. For example, generating checksums as part of the ingest process and adding this fixity information to the Archive Information Packages generated, or allowing manifests of checksums to be generated for multiple files and for the manifest and files to be bundled together for easy transport or storage. In addition md5sum and md5deep provide simple command line tools that operate across platforms to generate checksums on individual files or directories.

There are several different checksum algorithms, e.g. MD5 and SHA-256 that can be used to generate checksums of increasing strength. The ‘stronger’ the algorithm then the harder it is to deliberately change a file in a way that goes undetected. This can be important for applications where there is a need to demonstrate resistance to malicious corruption or alteration of digital materials, for example where evidential weight and legal admissibility is important. However, if checksums are being used to detect accidental loss or damage to files, for example due to a storage failure, then MD5 is sufficient and has the advantage of being well supported in tools and is quick to calculate.

The Handbook follows the National Digital Stewardship Alliance (NDSA) preservation levels (NDSA, 2013) in recommending four levels at which digital preservation can be supported through file fixity and data integrity techniques. Many of the benefits of fixity checking can only be achieved if there are multiple copies of the digital materials, for example allowing repair if integrity of one of the copies has been lost.

Level	Activity	Risks addressed and benefits achieved
1	• Check file fixity on ingest if it has been provided with the content. • Create fixity info if it wasn't provided with the content.	• Corrupted or incorrect digital materials are not knowingly stored. • Authenticity of the digital materials can be asserted. • Baseline fixity established so unwanted data changes have potential to be detected.
2	• Check fixity on all ingests • Use write-blockers when working with original media • Virus-check high risk content.	• No digital material of unconfirmed integrity can enter preservation storage. Evidential weight supported for authenticity. • Assurance can be given to all content providers that their content has been safely received. Original media is protected. • No malicious content can enter preservation storage.
3	• Check fixity of content held on preservation storage systems at regular intervals. • Maintain logs of fixity info and supply audit on demand. • Ability to detect corrupt data. • Virus-check all content.	• Protection from wide range of data corruption and loss events. Problems with storage are detected earlier. • Data corruption or loss does not go undetected due to 'silent errors' or 'undetected failures'. Digital materials are not in a state of 'unknown' integrity. • Ongoing evidential weight can be given that digital materials are intact and correct.
4	• Check fixity of all content in response to specific events or activities • Ability to replace/repair corrupted data • Ensure no one person has write access to all copies.	• Failure modes that threaten digital materials are proactively countered. All copies of digital materials are actively maintained. • Assurance to users of the integrity and authenticity of digital materials being accessed. • Effectiveness of preservation approach can be measured and demonstrated. • Compliance with standards, e.g. ISO 16363 Audit and certification of trustworthy digital repositories.

Write-blocking

Note that the National Digital Stewardship Alliance (NDSA) recommends the use of write-blockers at level 2. This is to prevent write access to media that digital materials might be on prior to being copied to the preservation storage system. For example, if digital material is delivered to an organisation on a hard disc drive or USB key then a write blocker would prevent accidental deletion of this digital material when the drive or key is read. Digital material might not be on physical media, e.g. it could be on a legacy storage server or delivered through a network transfer, e.g. an ftp upload. In these cases write blockers wouldn't apply and other measures would be used to make the digital material 'read only' on the source and hence immutable before confirmation that the digital material has been successfully transferred to preservation storage. Write blockers also don't exist for all types of media. If a write-blocker is applicable then the costs/skills required to use them should be balanced against the risk of damage to the original digital material or the need to have rigorous data authenticity. Therefore, some organisations might consider use of write blockers to be unnecessary or a level 3 or level 4 step.

Resources

Bailey, J., 2014, Protect Your Data: File Fixity and Data Integrity, The Signal, Library of Congress.

http://blogs.loc.gov/digitalpreservation/2014/04/protect-your-data-file-fixity-and-data-integrity/

Checking Your Digital Content: What is Fixity and When Should I Be Checking It?

http://digitalpreservation.gov/ndsa/working_groups/documents/NDSA-Fixity-Guidance-Report-final100214.pdf?loclr=blogsig

Many in the preservation community know they should be checking the fixity of their content, but how, when and how often? This document published by NDSA in 2014 aims to help stewards answer these questions in a way that makes sense for their organization based on their needs and resources (7 pages).

AVPreserve Fixity Tool

http://www.avpreserve.com/tools/fixity/

MD5

https://tools.ietf.org/html/rfc1321

SHA-1

http://csrc.nist.gov/publications/fips/fips180-4/fips-180-4.pdf

SHA-256

http://csrc.nist.gov/groups/STM/cavp/documents/shs/sha256-384-512.pdf

Md5deep and hashdeep

http://coptr.digipres.org/Md5deep_and_hashdeep

md5sum

http://coptr.digipres.org/Md5sum_Unix_command

The "Checksum" and the Digital Preservation of Oral History

https://www.youtube.com/watch?v=Emom_ncMqu0

A good short overview not limited to oral history, this video provides a brief introduction to the role of the checksum in digital preservation. It features Doug Boyd, Director of the Louie B. Nunn Center for Oral History at the University of Kentucky Libraries. (3 mins 25 secs)

References

Bailey, J., 2014. Protect Your Data: File Fixity and Data Integrity.*The Signal*. [blog]. Available: http://blogs.loc.gov/digitalpreservation/2014/04/protect-your-data-file-fixity-and-data-integrity/

ISO, 2012. ISO 14721:2012 - *Space Data and Information Transfer Systems – Open Archival Information System (OAIS) – Reference Model, 2nd edn*. Geneva: International Organization for Standardization. Available: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=57284

NDSA , 2013. *The NDSA Levels of Digital Preservation: An Explanation and Uses, version 1 2013*. National Digital Stewardship Alliance. Available: http://www.digitalpreservation.gov/ndsa/working_groups/documents/NDSA_Levels_Archiving_2013.pdf

File formats and standards

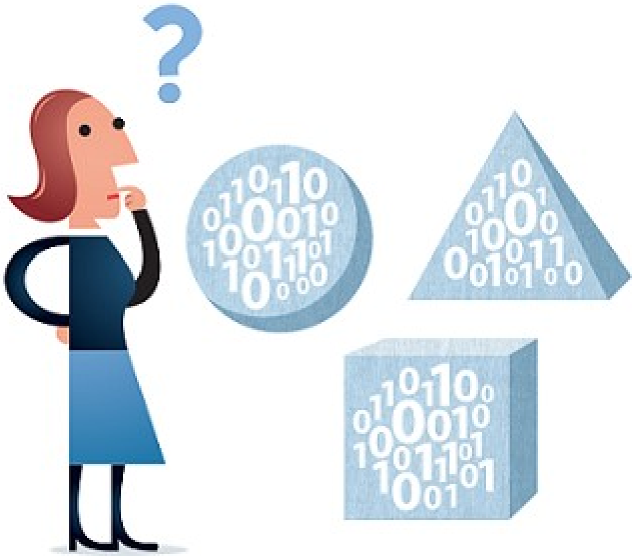

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Introduction

The management of file formats should be considered in the wider strategic context of preservation planning. What can your organisation afford to do? How much developer effort will it require? What do the users require from your collections? Are you committing yourself to a storage problem? At all times, the answer to digital preservation issues is not to try and “do everything”. Your strategy ought to move you towards simple and practical actions, rather than trying to support more file formats than you need.

The purpose of this section is not to provide a detailed or exhaustive list of current formats for different types of content but to draw attention to the broader implications of file formats for their application, and implications for preservation.

A substantial part of this chapter refers to the possible selection of a file format for migration purposes. While migration is a valid preservation strategy, and quite common for many file formats, it is not the only approach or solution. Where appropriate, the chapter will refer to other suitable methods for preservation.

File formats organised by content types

Different content types have, over time, developed their own file formats as they strive to accommodate functionality specific to their needs. The main content types are images, video, audio and text; however, a growing number of formats are being structured to address the demands of new media, including formats for 3D models and archiving the web.

File formats vary enormously in terms of complexity, with some data being encoded in many layers. In some cases the file formats involved are just one part of a larger picture, a picture that includes software, hardware, and even entire information environments.

For further advice on preservation of specific types of digital content and associated file formats see the Content-specific preservation case studies in the Handbook.

File formats - what should we be worrying about?

Obsolescence

Formats evolve as users and developers identify and incorporate new functionality. New formats, or versions of formats, may introduce file format obsolescence as newer generations of software phase out support for older formats. When software does not provide for backwards compatibility with older file formats, data may become unusable. Both open source and commercial formats are vulnerable to obsolescence: vendors sometimes use planned obsolescence to entice customers to upgrade to new products while open source software communities may withdraw support for older formats if these are no longer generally needed by the community. Obsolescence can also be accidental: both businesses and open source communities can fail.

File format format obsolescence is a risk that needs to be understood. That said, the problem may not be as severe as the digital preservation community perceived it to be some 10 years ago. Many established file formats are still with us, still supported, and still usable. It is quite likely that the majority of file formats you deal with will be commonly understood and well supported.

Proliferation

Arguably, in some sectors, proliferation is more of a challenge than obsolescence. If formats aren’t normalised then an organisation can end up with a large number of different file formats, and versions of those formats: e.g. lots of different versions of PDF, word, image formats etc. In domains which develop rapidly evolving bespoke data formats this problem can be exacerbated. Tracking and managing all these formats - which ones are at risk, and which tools can be used for each one - can be a serious challenge.

Your digital preservation strategy should strive to mitigate the effects of obsolescence and proliferation. Strategies as migration, emulation, normalisation and a careful selection of file formats are all valid and worth considering, in the context of your collections and your organisation.

Aspects of file formats for digital preservation

Selecting target formats for preservation

Not all digital formats are suited or indeed designed for archiving or preservation. Any preservation policy should therefore recognise the requirements of the collection content and decide upon a file format which best preserves those qualities. Pairing content with a suitable choice of preservation format or access format; identifying what is important in the content.

Below we suggest some factors to consider in selecting your preferred file formats:

Open source vs proprietary?

Open source formats, such as JPEG2000, are very popular due to their non-proprietary nature and the sense of ownership that stakeholders can attain with their use. However, the choice of open source versus proprietary formats is not that simple and needs to be looked at closely. Proprietary formats, such as TIFF, are seen as being very robust; however, these formats will ultimately be susceptible to upgrade issues and obsolescence if the owner goes out of business or develops a new alternative. Similarly, open source formats can be seen as technologically neutral, being non-reliant on business models for their development however they can also been seen as vulnerable to the susceptibilities of the communities that support them.

Although such non-proprietary formats can be selected for many resource types this is not universally the case. For many new areas and applications, e.g. Geographical Information Systems or Virtual Reality only proprietary formats are available. In such cases a crucial factor will be the export formats supported to allow data to be moved out of (or into) these proprietary environments.

Documentation and standards

The availability of documentation - for example, published specifications - is an important factor in selecting a file format. Documentation may exist in the form of vendor's specifications, an international standard, or may be created and maintained within the context of a user community. Look for a standard which is well-documented and widely implemented. Make sure the standard is listed in the PRONOM file format registry.

Adoption

A file format which is relied upon by a large user group creates many more options for its users. It is worth bearing in mind levels of use and support for formats in the wider world, but also finding out what organisations similar to you are doing and sharing best practice in the selection of formats. Wide adoption of a format can give you more confidence in your preservation strategy.

Lossless vs lossy

Lossy formats are those where data is compressed, or thrown away, as part of the encoding. The MP3 format is widely used for commercial distribution of music files over the web, because the lossy encoding process results in smaller file sizes.

TIFF is one example of an image format that is capable of supporting lossless data. It could hold a high-resolution image. JPEG is an example of a lossy image file format. Its versatility, and small file size, makes it a suitable choice for creating an access copy of an image of smaller size for transmission over a network. It would not be appropriate to store the JPEG image as both the access and archival format because of the irretrievable data loss this would involve.

One rule of thumb could be to choose lossless formats for the creation and storage of "archival masters"; lossy formats should only be used for delivery / access purposes, and not considered to be archival. A rule like this is particularly suitable for a digitisation project, particularly still images.

Support for metadata

Some file formats have support for metadata. This means that some metadata can be inscribed directly into an instance of a file (for example, JPEG2000 supports some rights metadata fields). This can be a consideration, depending on your approach to metadata management.

Significant properties of file formats

This is a complex area. One view regards significant properties as the "essence" of file content; a strategy that gets to the heart of "what to preserve". What does the user community expect from the rendition? What aspects of the original are you trying to preserve? This strategy could mean you don’t have to commit to preserving *all* aspects of a file format, only those that have the most meaning and value to the user.

Significant properties may also refer to a very specific range of *technical metadata* that is required to be present in order for a file to be rendered (e.g. image width). Some migration tools may strip out this metadata, or it may become lost through other curation actions in the repository. The preservation strategy needs to prevent this loss happening. It thus becomes important to identify, extract, store and preserve significant properties at early stage of the preservation process.

Things we can do

There are many things you could do to support file formats in your digital archive, and there are many tools available to help you with these tasks. There are now so many that digital preservation tool registries are being developed to help you locate and assess them (see the Tools and the Resources sections)

Tools for migration

Broadly, these are tools that transform a file format from an obsolete format into a newer format which can be supported. Many tools exist for doing this migration. They tend to confine themselves to doing one thing (e.g. ImageMagick only works for digital image objects).

A migration tool is just one part of a migration pathway. The pathway must include a destination / target format, which you will have selected in line with guidance as suggested above.

Migration tools may introduce risks. One of these risks is “invisible” changes happening to the content or to the data in the migration. To reduce this risk, one strategy is to devise a set of acceptance criteria for what the transformed object must keep, e.g. in terms of formatting, look and feel, or even functionality, and confirm desired outcomes with a process of quality assurance.

File format migration is not always the solution. Some CAD and CAM file formats cannot easily be migrated, for example. The aerospace industry has found that migration of older CAD files to a newer format requires a lot of validation, mainly because they are required by a regulatory framework to demonstrate that their data is sound and meets very strict standards. In short, the cost of migration and validation is (for them) much higher than an emulation solution, an approach which (in this case) involves keeping the CAD software and maintaining it.

See also the Tools and Content-specific preservation sections.

Tools for rendition

Broadly, these are tools that can read and play back a file format, so that the user community can read and interpret the resource; it's most commonly applied to files stored in accessible formats. A basic rendition tool would be PDF Reader. A more sophisticated rendition tool would be the Wellcome Library media player, which supports OCR texts, images, and audio-visual files.

Tools for file format identification

Tools that can identify aspects of file formats which are not immediately obvious from their file extension. They do this by reading the file format header, and thus can identify e.g. mimetype, size, version. Examples of such tools include PRONOM, JHOVE, and the NZ Metadata Extraction Tool (see Resources below).

These tools are usefully deployed at point of ingest, so that you know from the start what sort of file formats you are taking into the archive.

Some identification tools can also point to likely rendition tools, or even (like PRONOM) suggest a migration path based on file format identification.

Tools for file format validation

JHOVE is one of the few tools that is able to validate a file format. It does this by comparing an instance of a file format with sets of expected behaviours, which it stores in its library. JHOVE can report on certain file formats and tell whether they are valid and well-formed.

Collection surveys

Survey file formats in use / know what you have / characterisation of your collections. This again ties into a planning strategy, letting you know what you need to support, and the likely effort required to do this.

A survey should pay particular attention to *versions* of file formats, and software needed for their reading / rendition. If possible, gather any information about *published specifications* for these formats; some specs are published on the web.

Useful emerging work in this area has taken place at the British Library, with projects on Sustainability Assessments (Maureen Pennock, Paul Wheatley, Peter May) and Collection Profiling (Michael Day, Maureen Pennock, Ann MacDonald). At time of writing there are no active links to these projects, but it is anticipated that the Sustainability Assessment work will be published on the DPC wiki. These are useful approaches and can be regarded as examples of current best practice. Even if you don't assess or profile to the same depth as the BL, the exercise is a practical and applicable one.

Avoid Proliferation of File Types

Where possible, reduce the range of file formats you support, in order to reduce complexity. A sound approach to preservation planning is to normalise, rather than add multiple migration formats to your collection. The smaller the range of formats, the lower the overheads.

Community

Identify a consensus of agreement on target file formats; collaborate with institutions who hold similar collections to yours. What formats do they choose to work with?

Conclusion

For some kinds of content, there is consensus around the choice of preservation format. For example audio archiving where WAV is commonly used. In other areas consensus is much more difficult to achieve. The preservation of digital video is a complex area where progress has been stymied by a lack of agreement, and an uncontrolled proliferation of wrapper formats, delivery methods, and encoding methods. The choice of image file formats is slightly clearer, with a limited choice of formats for archiving and others for delivery. It has been generally agreed that the TIFF format is the correct format for archiving master files (the RAW or DNG format is also considered appropriate for archiving) but this is now being challenged by the JPEG2000 format which provides a far greater level of lossless compression compared to TIFF and is open source.

Resources

Library of Congress recommended format specifications

http://www.loc.gov/preservation/resources/rfs/index.html

develop a set of specifications of formats which it recommends, both internally to its own professionals and externally to creators, vendors and archivists, as the preferred ones to use to ensure the preservation and long-term access. It covers both digital and analogue formats and is divided into six broad categories: Textual Works and Musical Compositions; Still Image Works; Audio Works; Moving Image Works; Software and Electronic Gaming and Learning; and Datasets/Databases.

Jisc significant properties reports

Between 2007 and 2008 Jisc funded five studies of significant properties for different types of content and files. Note discussion in the reports is as of 2007- 2008. The reports are as follows:

inSPECT Significant Properties Report 2007 (10 pages)

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.109.7923&rep=rep1&type=pdf

Significant Properties of E-learning Objects 2008 (65 pages)

http://www.webarchive.org.uk/wayback/archive/20140616090345/http://www.jisc.ac.uk/media/documents/programmes/preservation/spelos_report.pdf

The Significant Properties of Moving images 2008 (62 pages)

http://www.webarchive.org.uk/wayback/archive/20140616090254/http://www.jisc.ac.uk/media/documents/programmes/preservation/spmovimages_report.pdf

The Significant Properties of Software: A Study 2008 (97 pages)

http://www.webarchive.org.uk/wayback/archive/20100624233431/http://www.jisc.ac.uk/media/documents/programmes/preservation/spsoftware_report_redacted.pdf

The Significant Properties of Vector Images 2007 (61 pages)

http://www.webarchive.org.uk/wayback/archive/20140616090304/http://www.jisc.ac.uk/media/documents/programmes/preservation/vector_images.pdf

British Library File Formats Assessments

http://wiki.dpconline.org/index.php?title=File_Formats_Assessments

The Digital Preservation Team at the British Library has undertaken preservation risk file format assessments to capture knowledge about the gaps in current best practice, understanding and capability in working with specific file formats. The focus of each assessment is on capturing evidence-based preservation risks and the implications of institutional obsolescence which lead to problems maintaining the content over time. The assessments are hosted as a new section on the DPC Wiki. Three assessments covering JP2, TIFF and PDF have commenced the series.

Library of Congress sustainability factors

http://www.digitalpreservation.gov/formats/index.shtml

This site is concerned with the formats associated with media-independent digital content, i.e., content that is typically managed as files and which is generally not dependent upon a particular physical medium. It is not concerned with the formats associated with media-dependent digital content, i.e., formats that are dependent upon and inextricably linked to physical media, e.g., DVDs, audio CDs, and videotape formats like DigiBeta. It identifies and describes the formats that are promising for long-term sustainability, and develops strategies for sustaining these formats including recommendations pertaining to the tools and documentation needed for their management.

Jisc digital media infokit: digital file formats

http://www.jiscdigitalmedia.ac.uk/infokit/file_formats/digital-file-formats

This Jisc Digital Media Infokit aims to provide quick and practical answers to 'what file format should I use for...? It covers still image, audio and video formats and common tasks and applications in education and heritage settings.

Help Solve the File Format Problem

http://fileformats.archiveteam.org

A crowd-sourced file format information wiki on the Archive Team site. All content is available under a Creative Commons 0 licence.

Is JPEG 2000 a digital preservation risk?

http://blogs.loc.gov/digitalpreservation/2013/01/is-jpeg-2000-a-preservation-risk/

An interesting guest blog and discussion thread on the JPEG 2000 image format.

OPF File Format Risk Registry

http://wiki.opf-labs.org/display/TR/OPF+File+Format+Risk+Registry

This focuses specifically on file format issues and risks that have implications for long-term preservation and accessibility and how to deal with these in a practical way. It aims to be complementary to more formal format registries.

PRONOM

http://apps.nationalarchives.gov.uk/pronom/Default.aspx

This file format registry is a major resource for anyone requiring impartial and definitive information about the file formats, software products and other technical components required to support long-term access to electronic records and other digital objects of cultural, historical or business value.

DROID (Digital Record Object Identification)

http://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/

This is an automatic file format identification tool providing categories of format identification for unknown files in a digital collection. It uses internal signatures to identify and report the specific file format and version of digital files. These signatures are stored in an XML signature file, generated from information recorded in the PRONOM registry.

Case studies

See the Detailed content preservation case studies section of the Handbook for relevant case studies.

Information security

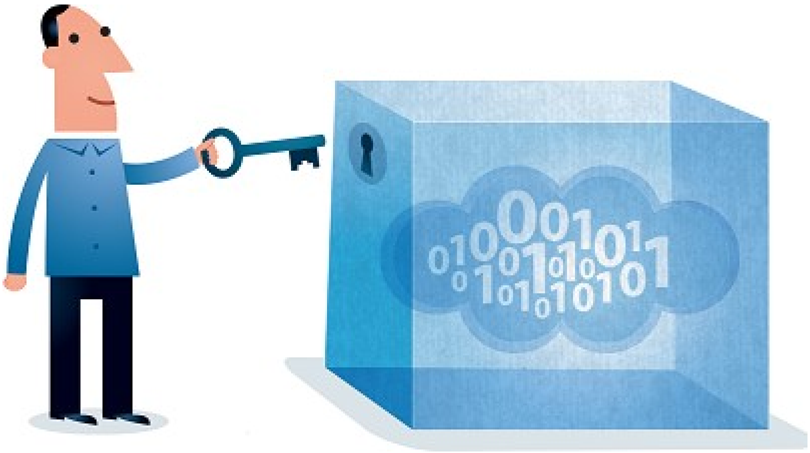

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Introduction

This section is intended as guidance for practitioners at a novice or intermediate level on the implications of information security for digital preservation. Information Security issues relate to system security (e.g., protecting digital preservation and networked systems / services from exposure to external / internal threats); collection security (e.g., protecting content from loss or change, the authorisation and audit of repository processes); and the legal and regulatory aspects (e.g. personal or confidential information in the digital material, secure access, redaction). Information security is a complex and important topic for information systems generally. It is important to rely on relevant expertise within your organisation and beyond it through government and other networks for general information security procedures and advice. You may also need appropriate advocacy for specific digital preservation procedures and requirements.

Rigorous security procedures will:

1. Ensure compliance with any legal and regulatory requirements;
2. Protect digital materials from inadvertent or deliberate changes;
3. Provide an audit trail to satisfy accountability requirements;
4. Act as a deterrent to potential internal security breaches;
5. Protect the authenticity of digital materials;
6. Safeguard against theft or loss.

Many types of digital material selected for long-term preservation may contain confidential and sensitive information that must be protected to ensure they are not accessed by non-authorised users. In many cases these may be legal or regulatory obligations on the organisation. These materials must be managed in accordance with the organisation's Information Security Policy to protect against security breaches. ISO 27001 describes the manner in which security procedures can be codified and monitored (ISO, 2013a). ISO 27002 provides guidelines on the implementation of ISO 27001-compliant security procedures (ISO, 2013b). Conforming organisations can be externally accredited and validated. In some cases your own organisation's Information Security Policy may also impact on digital preservation activities and you may need to enlist the support of your Information Governance and ICT teams to facilitate your processes.

Information security methods such as encryption add to the complexity of the preservation process and should be avoided if possible for archival copies. Other security approaches may therefore need to be more rigorously applied for sensitive unencrypted files; these might include restricting access to locked-down terminals in controlled locations (secure rooms), or strong user authentication requirements for remote access. However, these alternative approaches may not always be sufficient or feasible. Encryption may also be present on files that are received on ingest from a depositor, so it is important to be aware of information security options such as encryption, the management of encryption keys, and their implications for digital preservation.

Techniques for protecting information

Several information security techniques may be applied to protect digital material:

Encryption

Encryption is a cryptographic technique which protects digital material by converting it into a scrambled form. Encryption may be applied at many levels, from a single file to an entire disk. Many encryption algorithms exist, each of which scramble information in a different way. These require the use of a key to unscramble the data and convert it back to its original form. The strength of the encryption method is influenced by the key size. For example, 256-bit encryption will be more secure than 128-bit encryption.

It should be noted that encryption is only effective when a third party does not have access to the encryption key in use. A user who has entered the password for an encrypted drive and left their machine powered on and unattended will provide third parties with an opportunity to access data held in the encrypted area, which may result in its release.

Similarly encryption security measures (if used) can lose their effectiveness over time in a repository: there is effectively an arms race between encryption techniques and computational methods to break them. Hence, if used, all encryption by a repository must be actively managed and updated over time to remain secure.

Encrypted digital material can only be accessed over time in a repository if the organisation manages its keys. The loss or destruction of these keys will result in data becoming inaccessible.

Access Control

Access controls allow an administrator to specify who is allowed to access digital material and the type of access that is permitted (for example read only, write). The Handbook follows the National Digital Stewardship Alliance (NDSA) preservation levels in recommending four levels at which digital preservation can be supported through access control. The NDSA levels focus primarily on understanding who has access to content, who can perform what actions on that content and enforcing these access restrictions (NDSA, 2013) as follows:

NDSA level	Activity
1	• Identify who has read, write, move and delete authorisation to individual files • Restrict who has those authorisations to individual files
2	• Document access restrictions for content
3	• Maintain logs of who performed what actions on files, including deletions and preservation actions
4	• Perform audit of logs

Redaction

Redaction refers to the process of analysing a digital resource, identifying confidential or sensitive information, and removing or replacing it. Common techniques applied include anonymisation and pseudonymisation to remove personally identifiable information, as well as cleaning of authorship information. When related to datasets this is usually carried out by the removal of information while retaining the structure of the record in the version being released. You should always carry out redaction on a copy of the original, never on the original itself.

The majority of digital materials created using office systems, such as Microsoft Office, are stored in proprietary, binary-encoded formats. Binary formats may contain significant information which is not displayed, and its presence may therefore not be apparent. They may incorporate change histories, audit trails, or embedded metadata, by means of which deleted information can be recovered or simple redaction processes otherwise circumvented. Digital materials may be redacted through a combination of information deletion and conversion to a different format. Certain formats, such as plain ASCII text files, contain displayable information only. Conversion to this format will therefore eliminate any information that may be hidden in non-displayable portions of a bit stream.

Resources

ENISA. 2013, Cloud Security Incident Reporting

https://www.enisa.europa.eu/activities/Resilience-and-CIIP/cloud-computing/incident-reporting-for-cloud-computing/

The EU's Agency for Network & Information Security offers recommendations on the ways in which cloud providers and their customers should respond to – and report – security breaches. (38 pages).

ISO 27001:2013, Information technology— Security techniques — Information security management systems — Requirements. Geneva: International Organization for Standardization

http://www.iso.org/iso/catalogue_detail?csnumber=54534

ISO 27001 describes the manner in which security procedures can be codified and monitored. Conforming organisations can be externally accredited and validated. A template for a set of policies aligned with the standard is available. Note that these are headings, to assist with policy creation, rather than policy statements. However, similar policy sets are in use in a substantial number of organisations. (23 pages).

ISO 27002:2013, Information technology – Security techniques – Code of practice for information security controls. Geneva: International Organization for Standardization

http://www.iso.org/iso/catalogue_detail?csnumber=54533

ISO 27002 provides guidelines on the implementation of ISO 27001-compliant security procedures. (80 pages)

ISO 27799:2008, Health informatics – Information security management in health using ISO/IEC 27002. Geneva: International Organization for Standardization

http://www.iso.org/iso/catalogue_detail?csnumber=41298

ISO 27799 provides specific advice on implementing ISO 27002 and 27001 in the healthcare sector. (58 pages)

Cabinet Office, 2009, HMG IA Standard No. 1 – Technical Risk Assessment

http://www.cesg.gov.uk/publications/Documents/is1_risk_assessment.pdf

A detailed discussion and standard intended for UK Risk Managers and Information Assurance Practitioners who are responsible for identifying, assessing and treating the technical risks to systems and services that handle, store and process digital government information. (114 pages).

Redaction toolkit (TNA 2011)

http://www.nationalarchives.gov.uk/documents/information-management/redaction_toolkit.pdf

This TNA toolkit was produced in 2011 to provide guidance on editing exempt material from information held by public bodies. It covers generic principles records in any media but has a small section specifically on electronic records and detailed guidance on methods for securely redacting electronic records of all types. (21 pages).

BitCurator

http://wiki.bitcurator.net/index.php?title=Main_Page

BitCurator is a suite of open source digital forensics and data analysis tools to help collecting institutions holding born-digital materials. Parts of the toolset help locate private and sensitive information on digital media and prepare materials for public access.

Information Commissioners Office (ICO): Information security (Principle 7)

https://ico.org.uk/for-organisations/guide-to-data-protection/principle-7-security/

The ICO website has guidance on reporting of security breaches and use of IT. For those working in organisations falling under the ICO's jurisdiction an understanding of what this guidance recommends is essential to starting conversations with ICT and Information Governance Colleagues as they will need to be assured that work can be carried out in compliance with ICO recommendations.

Access to the Secure Lab

http://ukdataservice.ac.uk/get-data/how-to-access/accesssecurelab

A number of confidential and sensitive microdata sources are becoming available through datalabs across the UK. These data are deemed potentially identifiable, and can only be accessed through a datalab facility (as opposed to download). In addition, researchers are asked to fullfil a number of additional application requirements. Some of these data may be accessed via the Secure Lab of the UK Data Service and this page provides useful overviews and access to relevant user agreements.

Case studies

Opening access to administrative data for evaluating public services: The case of the Justice Data Lab

http://evi.sagepub.com/content/21/2/232.full.pdf+html

The Justice Data Lab a unit within a secure setting holding evaluation and statistical expertise has enabled providers of programmes aimed at reducing re-offending to obtain evidence on how the impact of their interventions differs from that of a matched comparison group. This article explores the development of the Justice Data Lab, the methodological and other challenges faced, and the experiences of user organizations. The article draws out implications for future development of Data Labs and the use of administrative data for the evaluation of public services. (16 pages).

UK Data Service: Data Security

http://ukdataservice.ac.uk/manage-data/store/security.aspx

This webpage summarises how the UK Data Archive manages data security for its holdings. Data security may be needed to protect intellectual property rights, commercial interests, or to keep sensitive information safe. Arrangements need to be proportionate to the nature of the data and the risks involved. Attention to security is also needed when data are to be destroyed.

References

NDSA, 2013. *The NDSA Levels of Digital Preservation: An Explanation and Use*s, *version 1* (2013). Available:

http://www.digitalpreservation.gov/ndsa/working_groups/documents/NDSA_Levels_Archiving_2013.pdf

ISO, 2013a. *ISO 27001:2013 - Information technology - Security techniques - Information security management systems - Requirements*. Geneva: International Organization for Standardization. Available: http://www.iso.org/iso/catalogue_detail?csnumber=54534

ISO, 2013b. *ISO 27002:2013 - Information technology – Security techniques – Code of practice for information security controls*. Geneva: International Organization for Standardization. Available: http://www.iso.org/iso/catalogue_detail?csnumber=54533

Cloud services

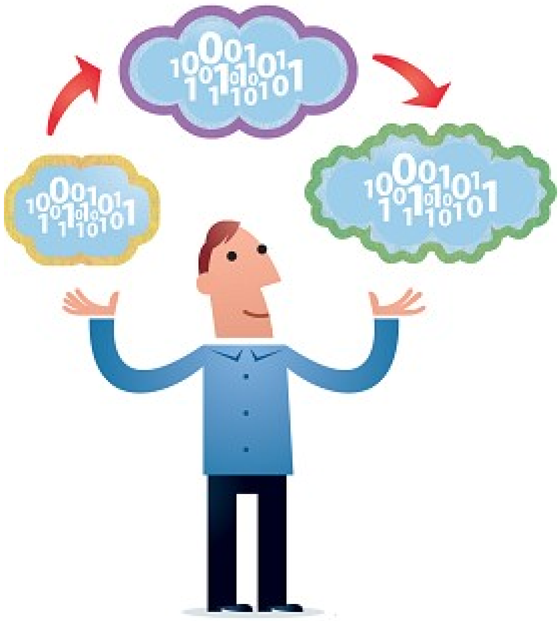

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

What is cloud computing?

Cloud Computing is a term that encompasses a wide range of use cases and implementation models. In essence, a computing ‘cloud’ is a large shared pool of computing resources including data storage. When someone needs additional computing power, they are simply able to check this out of the pool without much (often any) manual effort on the part of the IT team, which reduces costs and significantly shortens the time needed to start using new computing resources. Most of these ‘clouds’ are run on the public Internet by well-known companies like Amazon and Google. Some larger organisations have also found value in running private clouds inside their own data centres, where similar economies of scale begin to apply.

The generally accepted characteristics of a typical cloud service may be defined as computers and data storage which are:

- Available when required (‘on demand’), without the need for lengthy procurement and configuration processes;

- Available on standard networks such as the Internet, without special requirements for obscure or proprietary networking, protocols, or hardware;
- Able to offer additional capacity as demand increases, and less as demand falls ('elastic');
- Capable of only billing customers for the storage they use.

Cloud computing and digital preservation

Cloud computing can offer several benefits:

- The flexibility of the cloud allows relatively rapid and low-cost testing and piloting of emerging service providers. There are already some pilot activities with these cloud services and opportunities for shared learning across the community;
- There is now much greater flexibility and more options in deployment of cloud storage services and therefore greater relevance to archival repositories compared to earlier years (see Public, Community, Private and Hybrid clouds);
- There are potential cost savings from easier procurement and economies of scale, particularly for smaller repositories. These are important at a time of financial pressures;
- Cloud services can provide easy, automated replication to multiple locations essential for business recovery planning and access to professionally managed digital storage; in addition, the specialists can add access to other dedicated tools, procedures, workflow and service agreements, tailored for digital preservation requirements.

Cloud service models and service providers

There are four different cloud service models:

- Public – Commercial services hosted in large data centres around the world, accessible over public networks to anyone with the means to pay.
- Private - Large organisations create their own cloud by virtualising large sets of physical servers inside their own data centres.
- Hybrid – Combines aspects of combine aspects of public and private cloud , typically to handle large fluctuations in demand, or to satisfy different security requirements.
- Community - Architecturally, it may be effectively the same as a public cloud service, but optimised for a particular group of users to which access is restricted.

There are currently two classes of cloud service provider: generalists offering cloud storage (Amazon, Rackspace, Google, etc), and specialist companies that address additional specific digital preservation requirements and functions (see Resources and case studies for examples).

Positives

- Cloud services can provide easy, automated replication to multiple locations and access to professionally managed digital storage and integrity checking. As a result bit preservation (durability) of digital information can be at least as good (or better) than can be achieved locally;
- Archives can add access to dedicated tools, procedures, workflow and service agreements, tailored for digital preservation requirements via specialist vendors;

- There are potential cost savings from easier procurement and economies of scale, particularly for smaller archives;
- The flexibility of the cloud allows relatively rapid and low-cost testing and piloting of providers;
- There is much greater flexibility and more options in deployment of cloud services and therefore greater relevance to archives compared to earlier years. In particular private cloud or hybrid cloud implementations can address security concerns over storage of more sensitive material perhaps considered unsuitable for public cloud;
- Exit strategies can be put in place to address archival concerns over provider stability and longevity or other change risks. For example synchronising content across two cloud service providers or an external cloud with local internal storage; or agreeing an escrow copy held independently by a trusted third-party;
- There are already some pilot activities with these cloud services and opportunities for shared learning across the community.

Negatives

- The Cloud is designed for flexibility and rapid change. Archives however are long-term. Cloud storage and service contracts need careful management through time to meet archive needs. Data held in archives must be expected to be both preserved and accessible beyond the commercial lifespan of any current technology or service provider;
- Cloud can be cheaper, but it often requires organisations to think differently about the way their budgets are managed. There are also different skills to IT service vendor and contract management that may involve re-training or recruitment costs;
- Public cloud services tend to bill each month for capacity that has actually been consumed. As a result it can be difficult to budget ahead, or to accurately predict the amount of data likely to be uploaded, stored, or downloaded (however some vendors can invoice you for an annual subscription based on volume);
- As with any form of outsourcing, it is important that archives exercise due diligence in assessing and controlling the risks of cloud storage. Ensure that any legal requirements and obligations relating to third party rights in, or over, the data to be stored will be met. These may relate to management, preservation or access, and may have been placed upon archives and their parent organisations by their donors and funders via contracts and agreements or via legislation by Government;
- Use of cloud services will require archives to consider copyright-related questions including: who currently owns the copyright; whether additional licence permissions may be required; what permissions the cloud provider will need to provide the service; whether the cloud provider is able to use the data for their own purposes; and which party will own the rights in any data or works created from the original data;
- Use of cloud services may raise data security issues, where the relevant data is ‘personal data’ (e.g. data that permits the identification of a living individual), these include determining responsibility for securing data and audit of providers, as well as about location of processing and the extent to which risks incurred by automation of service provision can be addressed by contract;

- The legal elements of the relationship between an archive and a cloud service provider or providers (e.g. terms of service contracts and service level agreements) must be well defined and meet your requirements. This can be challenging as many cloud providers have standard SLAs and contracts to achieve commodity pricing and have limited flexibility on negotiating terms;
- Explicit provision must be made for pre-defined exit strategies and effective testing, monitoring and audit procedures.

Conclusions

The term "cloud" can encompass a wide range of implementation models for digital preservation services. There is much that can be learnt from organisations who have already piloted or moved to use of the cloud. For example several archives have been able to address the most widely held concerns over cloud services and find ways to successfully integrate cloud storage into their digital preservation activities. Others are using cloud based services for all or part of their other digital preservation functions such as preservation planning. Ultimately, procuring cloud services is similar to procuring any IT. You have to manage and address risks like you would for any other part of your IT infrastructure.

Resources

The National Archives Guidance on Cloud Storage and Digital Preservation (2nd Edition 2015)

http://www.nationalarchives.gov.uk/documents/CloudStorage-Guidance_March-2015.pdf

This guidance explores how cloud storage in digital preservation is developing, emerging options and good practice, together with requirements and standards that archives should consider. Sections focussing on services, legal issues, and five linked case studies, are provided. Sources of further advice and guidance are also included. (39 pages).

Aitken, B, McCann, P, McHugh, A and Miller, K, 2012, Digital Curation and the Cloud, DCC

http://www.jisc.ac.uk/media/7/C/1/%7B7C1A1FD7-44B4-4951-85A8-FC2C4CEB1564%7DCuration-in-the-Cloud_master_final.pdf

This 2012 report focused on the use of cloud services for research data curation. It provides some definitions of Cloud computing and examined a number of cloud approaches open to HE institutions in 2012. (30 pages).

Anderson. S, 2014, Feet On The Ground: A Practical Approach To The Cloud Nine Things To Consider When Assessing Cloud Storage, AV Preserve

http://www.avpreserve.com/wp-content/uploads/2014/02/AssessingCloudStorage.pdf

A white paper on cloud services, divided into nine topics and questions to ask. Vendor profiles against these nine topics are available. (7 pages).

A. Brown, C. Fryer, 'Achieving Sustainable Digital Preservation in the Cloud'

http://www.girona.cat/web/ica2014/ponents/textos/id87.pdf

This paper describes how Parliament is using the cloud as part of its digital repository infrastructure. 2004 (10 pages).

Digital Preservation Specialist Cloud Service Providers

ArchivesDirect

http://archivesdirect.org

ArchivesDirect features a hosted instance of Archivematica with storage via DuraCloud in secure, replicated Amazon S3 and Amazon Glacier storage.

Arkivum

http://arkivum.com

Arkivum's Archive as a Service provides a fully-managed and secure service for long-term data retention with online access and a guarantee of data integrity that's part of its Service Level Agreement and backed by worldwide insurance.

DuraCloud

http://www.duracloud.org

DuraCloud is a managed service from DuraSpace. It provides support and tools that automatically copies content onto several different cloud storage providers and ensures that all copies of the content remain synchronized. See also ArchivesDirect for its joint service with Archivematica.

Preservica

http://preservica.com/edition/cloud-edition/

Preservica Cloud Edition is a fully cloud hosted OAIS compliant digital preservation platform that also includes public access/discovery to allow you to safely share your archive or collection

David Rosethal's blog

http://blog.dshr.org/

Contains a number of posts on the economics of cloud computing

Case studies

The National Archives case study: Archives & Records Council Wales Digital Preservation Working Group

http://www.nationalarchives.gov.uk/documents/Cloud-Storage-casestudy_Wales_2015.pdf

This case study discusses the experience of a cross-sectoral working group of Welsh archives cooperating to test a range of systems and service deployments in a proof of concept for cloud archiving. It explains the organisational context, the varied nature of their digital preservation requirements and approaches, and their experience with selecting, deploying and testing digital preservation in the cloud. The case study examined the open source Archivematica software with Microsoft's Windows Azure; Archivematica with CloudSigma; Preservica Cloud Edition and has begun testing Archivematica with Arkivum 100. January 2015 (10 pages).

The National Archives case study: Tate Gallery

http://www.nationalarchives.gov.uk/documents/Cloud-Storage-casestudy_Tate_Gallery_2015.pdf

This case study discusses the experience of developing a shared digital archive for the Tate's four physical locations powered by a commercial storage system from Arkivum. It explains the organisational context, the nature of their digital preservation requirements and approaches, and their rationale for selecting Arkivum's on-premise solution, "Arkivum/OnSite" in preference to any cloud-based offerings. It concludes with the key lessons learned, and discusses plans for future development. January 2015 (7 pages).

The National Archives case study: Dorset History Centre

http://www.nationalarchives.gov.uk/documents/Cloud-Storage-case-study_Dorset_2015_%281%29.pdf

This case study covers the Dorset History Centre, a local government archive service. It explains the organisational context of the archive, the nature of its digital preservation requirements and approaches, its two year pilot project using Preservica Cloud Edition (a cloud-based digital preservation service), the archive's technical infrastructure, and the business case and funding for the pilot. It concludes with the key lessons they have learnt and future plans. January 2015 (9 pages).

The National Archives case study: Parliamentary Archives

http://www.nationalarchives.gov.uk/documents/Cloud-Storage-casestudy_Parliament_2015.pdf

This case study covers the Parliamentary Archives and their experience of procuring via the G-Cloud framework. For extra resilience/an exit strategy they have selected two cloud service providers with different underlying storage infrastructures. This is an example of an archive using a hybrid set of storage solutions (part-public cloud and part-locally installed) for digital preservation as the archive has a locally installed preservation system (Preservica Enterprise Edition) which is integrated with cloud and local storage and is storing sensitive material locally, not in the cloud. January 2015 (6 pages).

The National Archives case study: Bodleian Library, University of Oxford

http://www.nationalarchives.gov.uk/documents/Cloud-storage-casestudy_Oxford_2015.pdf

This case study covers the Bodleian Library and the University of Oxford, and the provision of a "private cloud" local infrastructure for its digital collections including digitised books, images and multimedia, research data, and catalogues. It explains the organisational context, the nature of its digital preservation requirements and approaches, its storage services, technical infrastructure, and the business case and funding. It concludes with the key lessons they have learnt and future plans. January 2015 (6 pages).

King's College London Kindura Project

http://link.springer.com/article/10.1186%2F2192-113X-2-13

The Kindura project led by King's College London and funded by Jisc, sought to pilot the use of a hybrid cloud for research data management. It used DuraCloud to broker between storage or compute resources supplied by external cloud services, shared services, or in-house services. There is an earlier Jisc prepared case study but a more recent open-access article on the project is linked.

University of Illinois Archives 2011 evaluation of Archivematica

http://e-records.chrisprom.com/evaluating-open-source-digital-preservation-systems-a-case-study-2/

Angela Jordan describes a 2011 evaluation by the University of Illinois Archives of Archivematica—an open-source, OAIS Reference Model-compliant digital preservation system. Because Archivematica was then in its alpha stages, working with this system was a way to explore what the system offered in relation to the needs of the University Archives, as well as provide input to the developers as they continued to refine the software for production release.

Digital forensics

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Introduction

Digital forensics is associated in many people’s minds primarily with the investigation of wrongdoing. However, it has also emerged in recent years as a promising source of tools and approaches for facilitating digital preservation and curation, specifically for protecting and investigating evidence from the past.

Institutional repositories and professionals with responsibilities for personal archives and other digital collections can benefit from forensics in addressing digital authenticity, accountability and accessibility. Digital personal information must be handled with due sensitivity and security while demonstrably protecting its evidential value.

Forensic technology makes it possible to: identify privacy issues; establish a chain of custody for provenance; employ write protection for capture and transfer; and detect forgery or manipulation. It can extract and mine relevant metadata and content; enable efficient indexing and searching by curators; and facilitate audit control and granular access privileges. Advancing capabilities promise increasingly effective automation in the handling of ever higher volumes of personal digital information. With the right policies in place, the judicious use of forensic technologies will continue to offer theoretical models, practical solutions and analytical insights.

Forensics in practice

There are three basic and essential principles in digital forensics: that the evidence is acquired without altering it; that this is demonstrably so; and that analysis is conducted in an accountable and repeatable way. Digital forensic processes, hardware and software have been designed to ensure compliance with these requirements.

Information assurance is critical. Writeblockers ensure that information is captured without altering it, while chains of custody in terms of evidence handling, process control, information audit, digital signatures and watermarking protect the historical evidence from future alteration and uncertain provenance.

Selective redaction, anonymization and encryption, malware sandbox containment and other mechanisms for security and fine-tuned control are required to assure that privacy is fully protected and inadvertent information leakage is prevented. Family computers, portable devices and shareable cloud services all harbour considerable personal information and consequently raise issues of privacy. Digital archivists and forensic practitioners share the need to handle the ensuing personal information responsibly.

The current emphasis on automation in digital forensic research is of particular significance to the curation of cultural heritage, where this capability is increasingly essential in a digital universe that continues to expand exponentially. Current research is directed at handling large volumes efficiently and effectively using a variety of analytical techniques. Parallel processing, for example, through purpose-designed Graphics Processing Units (GPUs), and high performance computing can assist processor-intensive activities such as full search and indexing, filtering and hashing, secure deletion, mining, fusion and visualization.

Especially noteworthy for digital preservation and curation is the way that digital forensics directs attention towards the digital media item as a whole – typically the forensic disk image, the file that represents everything on the original disk.

Forensic technologies

Forensic technologies vary greatly in their capability, cost and complexity. Some equipment is expensive, but some is free. Some techniques are very straightforward to use, others have to be applied with great care and sophistication. The BitCurator Consortium has been an important development bringing together a community of archival users of open source digital forensic tools (Lee et al, 2014). There is an increasingly rich set of open source forensic tools that are free to obtain and use – most significantly for archivists, BitCurator. These are a wonderful introduction to the ins-and-outs of digital forensics, and can be used to compare and cross-check the outputs of commercial or other open source tools.

Digital archivists and forensic specialists share a common need to monitor and understand how technology is used to create, store, and manage digital information. Additionally, there is a mutual need to manage that information responsibly in conformance with relevant standards and best practice. New forensic techniques are furthering the handling of digital information from mobile devices, networks, live data on remote computers, flash media, virtual machines, cloud services, and encrypted sources. The use of encryption is beginning to present significant challenges for digital preservation. It is not only a matter of decryption but of identifying encryption in the first place. Digital forensics offers some solutions.

Forensic and archival methodology must retain the ability both to retrospectively interpret events represented on digital devices, and to react quickly to the changing digital landscape by the rapid institution of certifiable and responsible policies, procedures and facilities. The pace of change also has implications for ongoing training of curators and archivists, and there are digital forensics courses endorsed by archival, scholarly and preservation institutions.

Conclusion

In conclusion, there are some deep challenges ahead for cultural heritage and archives, but the forensic perspective is undoubtedly among the most promising sources of insights and solutions. Equally, digital forensics can benefit from the advances being made in the curation and preservation of digital information.

This brief overview has been based on short excerpts from The Digital Preservation Technology Watch Report on Digital Forensics and Preservation (John, 2012) with additional material kindly provided by Jeremy Leighton John, the author of the report. See Resources and case studies for further detailed guidance and exemplars.

Resources

Digital forensics and preservation DPC technology watch report

http://dx.doi.org/10.7207/twr12-03

This 2012 DPC report provides a broad overview of digital forensics, with some pointers to resources and tools that may benefit cultural heritage and specifically the curation of personal digital archives (60 pages).

Digital forensics and born-digital content in cultural heritage collections

http://www.clir.org/pubs/reports/pub149/pub149.pdf/view

This CLIR report introduces the field of digital forensics in the cultural heritage sector and explores some points of convergence between the interests of those charged with collecting and maintaining born-digital cultural heritage materials and those charged with collecting and maintaining legal evidence (93 pages).

Archivematica

https://www.archivematica.org/wiki/Main_Page

Archivematica is an open source digital preservation system and has addressed the ingest of forensic disk images as part of its workflows and toolset.

BitCurator

http://www.bitcurator.net

The website provides access to information on the BitCurator Consortium (BCC), projects, and tools. BitCurator has developed, packaged and documented open-source digital forensics tools to allow libraries, archives and museums to extract digital materials from removable media in ways that reflect the metadata and ensure the integrity of the materials, allowing users to make sense of materials and understand their context, and preventing inadvertent disclosure of sensitive data. The consortium is an independent, community-led membership association that serves as the host and center of administrative, user and community support for the BitCurator environment.

Forensics wiki

http://forensicswiki.org/wiki/Main_Page

The Forensics Wiki is a Creative Commons-licensed wiki devoted to information about digital forensics. It lists over 700 pages focused on the tools and techniques used by investigators, important papers and reports, people, and organizations involved.

The Invisible Photograph Part 2: Trapped: Andy Warhol's Amiga Experiments

http://www.nowseethis.org/invisiblephoto/posts/108

A team of computer scientists, archivists, artists, and curators teamed up to unearth Andy Warhol's lost digital works on a 30 year old Amiga Commodore computer (18 mins 52 secs)

The Invisible Photograph Part 3: Extraterrestrial: The Lunar Orbiter Image Recovery Project

http://www.nowseethis.org/invisiblephoto/posts/384

How the "techno archaeologists" of the Lunar Orbiter Image Recovery Project digitally recovered the first photographs of the moon taken by a set of unmanned space probes in the 1960s. (22 mins 07 secs)

Case studies

Carcanet email project

http://www.library.manchester.ac.uk/aboutus/projects/carcanet/

A Jisc-funded project that tackled the challenge of capturing and preserving the email archive of Carcanet Press, one of the UK's premier poetry publishing houses. It was winner of the 2014 DPC Preservation Wward for Safeguarding the Digital Legacy. The project explored and adopted several ediscovery and forensic tools, specifically AccessData's Forensic Toolkit (FTK), Paraben's Email Examiner and Fookes Software's Aid4Mail eDiscovery. A project final report summarizes the work (Baker, 2014).

References

John, J. L., 2012. Digital Forensics and Preservation. *DPC Technology Watch Report* 12-03 November 2012. Available: http://dx.doi.org/10.7207/twr12-03

Lee, C. A., Olsen, P., Chassanoff, A., Woods, K., Kirschenbaum, M. & Misra, S., 2014. *From Code to Community: Building and Sustaining BitCurator through Community Engagement.* BitCurator White Paper 30 September 2014. Available: http://www.bitcurator.net/wp-content/uploads/2014/11/code-to-community.pdf

Persistent identifiers

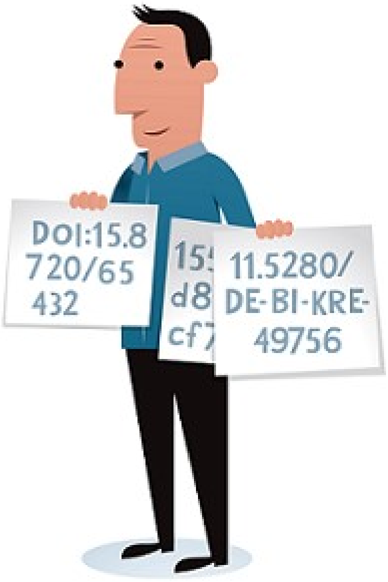

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Introduction

This section provides guidance on the use of persistent identifiers for digital objects and digital preservation. Other types of persistent identifier schemes exist e.g. for individuals or institutions.

A persistent identifier is a long-lasting reference to a digital resource. Typically it has two components: a unique identifier; and a service that locates the resource over time even when it's location changes. The first helps to ensure the provenance of a digital resource (that it is what it purports to be), whilst the second will ensure that the identifier resolves to the correct current location.

Persistent identifiers thus aim to solve the problem of the persistence of accessing cited resource, particularly in the academic literature. All too often, web addresses (links) fail to take you to the referenced resource you expect. This can be for technological reasons like server failure but human-created failures are more common. Organisations transfer journals to new publishers, reorganise their websites, or lose interest in older content, leading to broken links when you try to access a resource. This is frustrating for users, but the consequences can be serious if the linked resource is essential for legal, medical or scientific reasons.

Persistent identifiers can also be used 'behind-the-scenes' within a repository to manage some of the challenges in cataloguing and describing, or providing intellectual control and access to born-digital materials.

Schemes

Since the problem of persistence of an identifier is created by humans, the solution of persistent identifiers also has to involve people and services not just technologies. There are several persistent identifier schemes and all require a human service element to maintain their resolution systems. The main persistent identifier schemes currently in use are detailed below.

Digital Object Identifier (DOI)

DOIs are digital identifiers for objects (whether digital, physical or abstract) which can be assigned by organisations in membership of one of the DOI Registration Agencies; the two best known ones are CrossRef, for journal articles and some other scholarly publications, and DataCite for a wide range of data objects. As well as the object identifier, DOI has a system infrastructure to ensure a URL resolves to the correct location for that object.

Handle

Handles are unique and persistent identifiers for Internet resources, with a central registry to resolve URLs to the current location. Each Handle identifies a single resource, and the organisation which created or now maintains the resource. The Handle system also underpins the technical infrastructure of DOIs, which are a special type of Handles.

Archival Resource Key (ARK)

ARK is an identifier scheme conceived by the California Digital Library (CDL), aiming to identify objects in a persistent way. The scheme was designed on the basis that persistence "is purely a matter of service and is neither inherent in an object nor conferred on it by a particular naming syntax".

Persistent Uniform Resource Locator (PURL)

PURLs are URLs which redirect to the location of the requested web resource using standard HTTP status codes. A PURL is thus a permanent web address which contains the command to redirect to another page, one which can change over time.

Universal Resource Name (URN)

URNs are persistent, location-independent identifiers, allowing the simple mapping of namespaces into a single URN namespace. The existence of such a Uniform Resource Identifier does not imply availability of the identified resource, but such URIs are required to remain globally unique and persistent, even when the resource ceases to exist or becomes unavailable. The URN term is now deprecated except in the very narrow sense of a formal namespace for expressing a Uniform Resource Identifier.

Choosing a Persistent Identifier Scheme

There needs to be a social contract to maintain the persistence of the resolution service - either by the organisation hosting the digital resource, a trusted third party or a combination of the two. Each scheme has its own advantages and constraints but it is worth considering the following when deciding on a persistent identifier strategy or approach:

Advantages

- Critically important in helping to establish the authenticity of a resource.
- Provides access to a resource even if its location changes.
- Overcomes the problems caused by the impermanent nature of URLs.
- Allows interoperability between collections.

Disadvantages

- There is no single system accepted by all, though DOIs are very well established and widely deployed.
- There may be costs to establishing or using a resolver service.
- Dependence on ongoing maintenance of the permanent identifier system.

Conclusions

Persistent identifiers need to be supported by enduring services and are not just unique strings of alpha-numeric characters that are assigned to a digital resource. They have become particularly important for research data and e-journal articles (see content specific preservation section on e-Journals) and are a significant part of the long-term infrastructure for digital preservation of research. For the issue of link-rot for more general web pages, and solutions harnessing web-archives to resolve this see the content specific preservation section on Web-archiving.

Resources

Persistent identifiers - an overview. TWR Technology Watch Review

http://www.metadaten-twr.org/2010/10/13/persistent-identifiers-an-overview/

This article by Juha Hakala (2010) describes five persistent identifier systems (ARK, DOI, PURL, URN and XRI) and compares their functionality against the cool URIs. The aim is to provide an overview, not to give any kind of ranking of these systems.

Preservation, trust and continuing access for e-Journals DPC technology watch report

http://dx.doi.org/10.7207/twr13-04

This 2013 report by Neil Beagrie discusses current developments and issues which libraries, publishers, intermediaries and service providers are facing in the area of digital preservation, trust and continuing access for e-journals. It includes generic lessons and recommendations on outsourcing and trust of interest to the wider digital preservation community and covers relevant legal, economic and service issues as well as technology. (49 pages).

Persistent Identifiers in the Publication and Citation of Scientific Data

http://www.iza.org/en/papers/5090_28102009.pdf

Presentation by Jens Klump, German Research Centre for Geosciences (GFZ) on the DFG STD-DOI project, which details the background and reasoning behind the foundation of DataCite. 2009. (47 pages).

DCC Briefing Paper: Persistent Identifiers

http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation/persistent-identifiers

This 2006 paper by Joy Davidson discusses how progress in defining the nature and functional requirements for identifier systems is hindered by a lack of shared agreement on what identifiers should actually do; simply provide a globally or locally unique name for a digital or analogue resource, or incorporate associated services such as resolution and metadata binding. The application and maintenance of identifiers forms just one part of an overall digital preservation strategy; in order to offer any guarantees of persistence in the long or short-term they need institutional commitment and clearly defined roles and responsibilities. (2 pages)

ARK

http://www.cdlib.org/services/uc3/arkspec.pdf

CrossRef

http://www.crossref.org

DataCite

http://www.datacite.org

DOI

http://www.doi.org/

Handle

http://www.handle.net/

Perma.CC

https://perma.cc/about

PURL

https://purl.org/docs/index.html

URN

http://tools.ietf.org/html/rfc3986

Case studies

DCC case study: Assigning digital object identifiers to research data at the University of Bristol

http://www.dcc.ac.uk/resources/persistent-identifiers

The University of Bristol runs a dedicated research data repository as part of their Research Data Service. They are using the DataCite service at the British Library to assign digital object identifiers (DOIs) to research datasets in order to provide unique and perpetual identifiers for data, to allow easy citation and discoverability. The Bristol Research Data Service provides guidance on how to use the identifiers to cite data and is developing appropriate policies to monitor usage. 2004. (4 pages).

Links that Last

http://www.dpconline.org/events/previous-events/925-links-that-last

This DPC briefing day in July 2012 introduced the topics of persistent identifiers and linked data, and discussed the practical implications of both approaches to digital preservation. It considered the viability of services that offer persistent identifiers and what these offer in the context of preservation; reviewed recent developments in linked data, considering how such data sets might be preserved; and by introducing these two parallel topics it went on to consider whether both approaches can feasibly be linked to create a new class of robust linked data. A series of presentations including case studies are linked from the provisional programme.