Digital Preservation **Handbook**

Digital Preservation Briefing

Illustrations by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Who is it for?

Senior administrators (DigCurV Executive Lens), operational managers (DigCurV Manager Lens) and staff (DigCurV Practitioner Lens) within repositories, funding agencies, creators and publishers, anyone requiring an introduction to the subject.

Assumed level of knowledge

Novice.

Purpose

- To provide a strategic overview and senior management briefing, outlining the broad issues and the rationale for funding to be allocated to the tasks involved in preserving digital resources.
- To provide a synthesis of current thinking on digital preservation issues.
- To distinguish between the major categories of issues.
- To help clarify how various issues will impact on decisions at various stages of the life-cycle of digital materials.
- To provide a focus for further debate and discussion within organisations and with external audiences.

Reusing this information

Contents

Why Digital Preservation Matters

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Introduction

This section together with Preservation issues is designed as a briefing for those new to digital preservation. It is structured into four inter-linked sub-sections. In addition it has a close relationship to the Getting started section, which is also particularly designed with those new to digital preservation in mind.

Digital preservation: the challenge of a generation

Any digital object can be considered in scope for digital preservation: born digital or digitised, corporate or personal, innovative or routine. Digital preservation can encompass texts and images, databases and spreadsheets, vectors or rasters, programs and applications, desktop files and enterprise systems, email and social media, games, movies, music and sound, entire web domains and individual tweets. Digital collections can derive from laptops or desktops or smart phones; from tablets, souped-up servers or hulking great mainframes. They can be snapped at the end of a selfie stick or beamed from sensors deep in space; they can be generated by tills and cash machines, by satellites and scanners, by tiny sensitive chips and massive arrays. They can be stored in repositories or data centres or USB sticks. There is no digital object or system that is not provisionally within scope for digital preservation.

Pervasive, changing and ubiquitous, digital technologies are a defining feature of our age. Digital materials are a core commodity for industry, commerce and government. They are fundamental for research, the law and medicine. The creative industries, cultural heritage and the media depend on reliable access to digital materials while families and friends extend and sustain their relationships through digital interactions.

But digital materials - and the opportunities they create - are fragile even if they also have the capacity to be durable through replication. Digital platforms change and the long chains of interdependence on which they depend are complicated and fluid. Their longevity and utility is threatened where contents or contexts are lost: engagement and exploitation are enabled when digital materials endure. The greater the importance of digital materials, the greater the need for their preservation: digital

preservation protects investment, captures potential and transmits opportunities to future generations and our own.

Already we have made great strides in averting a "digital dark age". There are a growing number of repositories all over the world that can claim a long track record of keeping digital materials well over many decades (for example the UK Data Archive founded in 1967). This gives us a broad foundation of experience and collaborative professional networks to draw on.

It is a shared, generational challenge.

The always emerging digital preservation challenge

The unifying characteristic of digital materials is their machine-dependency. Information can only be accessed and functions can only be executed through a computer. As technology becomes more sophisticated this dependence becomes an ever more elaborate chain of inter-dependencies that are hard to track and tricky to maintain.

So long as the IT sector remains innovative in its provision of new tools and technologies, digital preservation managers will respond by devising effective strategies for ensuring the durability and usability of new digital materials, so digital preservation will remain an always-emerging challenge.

To ensure the value of digital materials in the long run we need to ensure access, which in turn means we need to understand and mitigate rapid changes in technology and organisations (see Preservation issues).

Digital material can often only be archived well in digital form: there is no non-digital equivalent such as paper that retains all the essential information and functionality it provides. Too often it has been necessary to print out digital material for archiving and then even re-digitizing the printed copy later because there has been no local capacity for managing born digital material.

Today we have a growing and effective body of approaches, experience, and collaboration to address the challenges. Digital preservation is an important, necessary and doable endeavour with simple first steps all can undertake (see Getting started).

What is in scope?

Simply because everything could be in scope for a digital preservation strategy does not mean that everything should be preserved.

The question is less what can be preserved so much as what should not be lost. Selection, appraisal and disposal are significant components in any digital management activity. In the context of an expanding digital universe, a determined effort to identify, process and retain digital material of enduring value means on one hand that the right material is available to the right people at the right time in the right format; and on the other hand material is identified that can be actively removed or benignly neglected.

Digital material provides profound new opportunities for access and use of repositories. If digital collections exist in a fast changing environment, then we should expect that our users do too. Users of digital materials are likely to be using technology that is not yet fully developed in ways that we cannot fully anticipate, in places we may never visit and for purposes that we may struggle to predict. So any meaningful answer to the question of 'how can we preserve digital materials' will rapidly resolve to 'what can we do to ensure that these digital materials can be used'? Preservation planning will only succeed when user needs are fulfilled.

All of this indicates a requirement that wherever possible the long term viability of digital materials should be defined early not late. Preservation action is needed at the start of the life of a digital object, not always at its end. Creation, management and archiving of digital materials are no longer at opposite ends of a process but are integrated all the way through. By extension, preservation is no longer simply a concern for memory institutions in the long term but for everyone interested in using and accessing digital materials.

Who needs to be involved?

The ability to preserve digital materials depends upon a wide range of stakeholders. Principal among these are the creators of digital content, whose involvement in their preservation might involve, for example, consideration of standards in terms of format and media, and ensuring enough contextual information is available to enable their management by others. Creators may often be unaware of their pivotal role. This could be for all kinds of reasons, but a vital part of any digital preservation effort is the effective dialogue with creators of digital materials to inform and advocate the value of their engagement (to them and others).

If the creators of digital materials have a responsibility to enable long term access, then this responsibility is borne even more fully by those who provide the infrastructure and environments in which they are created. In some cases this may be a corporate function, with the provision of corporate tools and services which are preservation ready. In other cases responsibility will be borne by external service providers who host digital infrastructure for clients.

The nature of digital technology dictates that it is not feasible simply to hand over stewardship of the resource at some point in the future, without having managed it sufficiently to facilitate sustainability.

In some cases, institutions will manage their own digital legacy: large institutions that create digital materials may most sensibly be the ones to manage them in the long term, thus maximising return on their initial investment. But in other contexts co-operative models for long-term preservation have emerged involving a number of organisations. Both subject specialist and expert centres have emerged offering specific preservation solutions for specific types of digital material.

For some organisations, it may prove more cost-effective to contract all or part of their digital preservation activities to a third party. Whilst it may be advantageous to outsource, it is important to remember responsibility remains with the organisation. Staff will need to be sufficiently aware of digital preservation issues, particularly as they relate to legal, organisational and contractual problems, to manage these third party contracts effectively.

Any institution which places value on digital resources in general needs to ensure the long-term preservation of digital materials. A significant number of institutions have not only taken that role on for themselves but have offered wider leadership in addressing the practical implications of digital preservation.

Ultimately however, digital preservation cannot be perceived as solely a concern for archives, libraries, museums and other memory institutions: it is a challenge for all who have an interest in creating, using, acquiring and making accessible, digital materials.

Resources

Why Digital Preservation is Important for Everyone

https://www.youtube.com/watch?v=qEmmeFFafUs&index=43&list=PLEA69BE43AA9F7E68

Short Library of Congress video produced in 2010 for the non-specialist audience explaining how traditional information sources such as books, photos and sculptures can easily survive for years, decades or even centuries but digital items are fragile and require special care to keep them useable. Rapid technological changes also affect digital preservation. As new technologies appear, older ones become obsolete, making it difficult to access older content. (2 mins 51 secs)

Preservation Issues

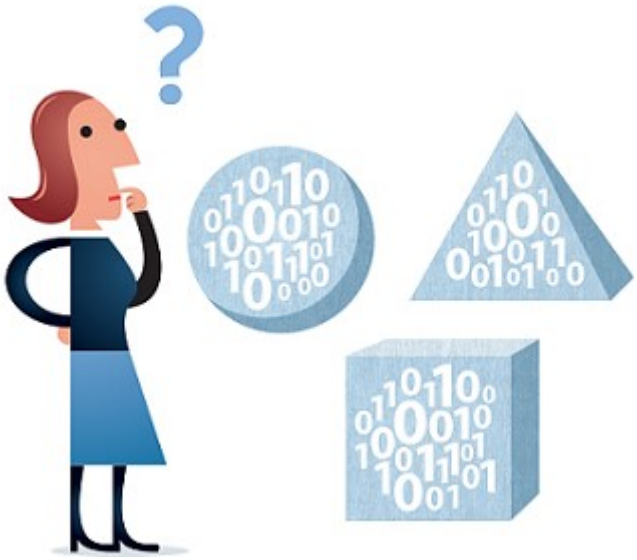
Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Introduction

This section together with Why digital preservation matters, is designed as a briefing for those new to digital preservation. It is structured into three inter-linked sub-sections covering Threats to digital materials, Organisational issues, and Resourcing issues. It links to more detailed treatment in other sections of the Handbook as appropriate, but has a particularly close relationship to the Getting started section, which is also particularly designed with those new to digital preservation in mind.

Digital preservation can often seem daunting at first. It is important to realise than those with existing skills in either information management or information technology within organisations are well placed to build on and apply these skills to digital preservation activities. However, it may require initially learning some new unfamiliar terminology (see Glossary), extending skill sets, and sometimes working in new ways.

Threats to Digital Materials



Keeping the data

Every digital file is formed from a series of zeros and ones, or bits (binary digits). These streams of bits need to be captured and retained over time, without loss or damage, to ensure the survival of digital materials. There are an array of threats to any attempt at preserving these bits. Storage media can decay over time, leading to corrupted files. Storage media may become obsolete and unsupported by contemporary computers and the software that understands and provides access to them. The bits may be ignored, abandoned, accidentally deleted or maliciously destroyed. Removable media could be left on a shelf and forgotten, files stored on a shared network drive might be left without an owner, or a third party cloud storage provider could go out of business.

Maintaining a systematic process for bit preservation remains a fundamental requirement in ensuring long term digital preservation. Storage media must be monitored and refreshed (See Legacy media). Redundancy must be introduced by replicating or backing up files, introducing diversity in dependent technologies and avoiding catastrophic disaster at a single geographical location (see Storage). Checksums must be generated and frequently recalculated to identify any loss and ensure that the integrity of the bits can be verified in an efficient and automated manner (see Fixity and checksums). The locations in which digital materials are stored should be carefully recorded, and responsibility for their preservation allocated.

Keeping the meaning of the data

Reconstructing the information that is encoded within a stream of a bits typically requires computer software that is designed to render, manipulate, analyse or otherwise interact with the particular encoding or format of the data. Over time, the encodings (or file formats) may change, and the software applications that interact with them may go in and out of favour. Although unusual for well known file formats, less well used file formats may become obsolete over time, as the software that renders them is no longer supported (see File formats and standards).

Understanding the technology on which particular digital materials are dependent enables appropriate action to be taken to ensure their preservation. A considered preservation planning process might result in the migration of digital files from format to format, the emulation of obsolete software, or the employment of alternative software applications to render the data (see Preservation

action). Each of the options presents its own advantages and disadvantages and these need to be evaluated carefully, possibly on a case by case basis (see Preservation planning).

While file format obsolescence has not emerged as the overwhelming danger that was previously perceived, challenging subtleties remain. It may be possible to find a method for rendering an old file format (perhaps by emulating some obsolete software), but how accurate is the rendering, is it legal to run the software, and how much will this complex effort cost the preserver and the user?

Maintaining trust in the data

Digital materials have the potential to remain fluid over time, being edited or altered with ease, being damaged by media failure, or decoded into human readable information in an unreliable or inaccurate manner by rendering software. For an end user to have trust in the result of digital preservation work it requires careful consideration of the entire lifecycle of the digital materials and who or what has interacted with them over time. Information management systems need to be able to link to essential contextual information regarding the business procedures of the creating agency. Authenticity and integrity of digital resources can be equally important in other sectors. For example, scholars will need to feel confident that references they cite will stay the same over time, courts of law will need to be assured that material can withstand legal evidential requirements, government departments may well have legally enforceable requirements regarding authenticity, and so on. This issue overlaps with both legal and organisational issues and it may be one which is best resolved within individual sectors rather than through generic procedures.

The application of data integrity techniques and the maintenance of audit trails can provide confidence that a digital object has remained unchanged (except by necessary preservation action) since deposit in an archive (see Fixity and checksums, and Information security). Ultimately its authenticity to a user may depend much more on the broader trustworthiness of the preserving organisation as a whole. Maintaining high quality preservation processes based on current best practice and validated by appropriate audit and certification will be crucial (see Audit and certification).

Keeping the context of the data and its dependencies

The meaning of digital information can be dependent on additional information that may have been implicit within the context it was originally created or used in, but less clear when revisited at a later date. Identifying, understanding and capturing relevant contextual information can be vital to a successful preservation effort. This might be as simple as capturing the units of measurement used within a spreadsheet, the scale of a map, or the point of origin within a CAD drawing. As digital information continues to be created in a more complex and interconnected manner, it may be necessary to retain the place of particular digital materials within a wider context of associated information resources. What may be seemingly simple and stand alone documents may actually depend on related files, referenced fonts and may have pointers to related information on the web. What might be viewed as a simple web page may have been generated on the fly from live data sourced from different locations on the Internet.

Understanding the data, how it will be used, its dependencies and its context will enable it to be captured for preservation in an appropriate manner and documented in a sufficiently explicit manner to enable the intellectual content to be retained and understood on into the future (see Metadata and documentation).

Acting in a timely manner

Prioritising digital preservation activities and applying them in a timely manner can be crucial not just in avoiding loss but in ensuring the best use of limited resources. Where the opportunity exists to intervene early in the lifecycle, digital materials can be shaped to survive better into the future. The choice of file format, the capture of critical documentation or the description of key relationships in the metadata may require a small investment up front, but could deliver considerable savings further down the line (see Creating digital materials). Where this is not possible, and risks to the data have been identified, the best timing for preservation action can be unclear. Early intervention to head off technological obsolescence may provide greater confidence of long term sustainability but with the risk that intervention may not ultimately be necessary and resources were wasted. Just in time action may minimise unnecessary activity, but increase the effort needed to research obsolete technology in a particular case requiring specialist knowledge that is no longer current. Appropriate action should be taken on a case by case basis.

Coping with the data deluge

Research reported by David Rosenthal noted that the rate of data creation is expanding by about 60% per annum; that developments in data storage allow are expanding at about 25% per annum; and that data centre budgets are expanding at about 2% per annum (Rosenthal, 2014). While this places challenging pressures on selection policies and other organisational decision making it also poses technological questions. Simple preservation processes that function effectively at one level will not necessarily scale easily to work with very large volumes of data or perhaps very large individual files. The technology and understanding to work at scale is moving forward rapidly, with growing expertise for handling large audio visual collections, research data and web based archives (see Content-specific preservation). But some repositories still face significant challenges in developing and maintaining scalable architectures and procedures to handle growing quantities of data. The technical and managerial challenges in accessioning, managing and providing access to digital materials on this scale should not be underestimated. It can be important to remember that selection, appraisal and disposal are significant components in any digital management activity.

Organisational Issues

While technological issues can be challenging, there are also numerous challenges which relate to organisational issues. These include how digital preservation is organised and delivered, or how those responsibilities change over both time and the lifecycle of digital materials. There are common digital preservation challenges faced across organisations, yet every organisational context will be different. It is vital to ascertain organisational drivers and tailor practical solutions to meet these needs. There is no one size fits all approach for digital preservation.

The creation, preservation and access for digital materials are widely distributed. As a result, there is an increasing need to go beyond the confines of individual organisations, or even countries, to maximise the benefits of the technology, address common issues, and to overcome the challenges cost-effectively.

In-house or outsource?

The decision whether to do all or part of digital preservation via a third-party or in-house, or perhaps a combination of the two, is often a complex one. Digital preservation may be undertaken in-house if there is sufficient staffing and infrastructure but outsourcing some activities or support can be cost-effective, and can leverage internal capabilities and capacity.

Outsourcing specific tasks or services from a repository is by no means a new phenomenon. Repositories have contracted out some of their operations for decades. Of critical importance is having and retaining sufficient knowledge to be able to prepare effective specifications and monitor performance. Outsourced work must be easily verified and quality checked, and this is best enabled via careful design of the specification, and the reporting providing by the 3rd party. Cost will clearly be a key consideration when deciding whether or not to contract out digital preservation but there are also other factors to consider such as legal issues. For example, legal provisions due to privacy or confidentiality may influence whether outsourcing is appropriate or not. The advantages and disadvantages of each option will need to be balanced in light of the individual organisation's mission and responsibilities (see Procurement and third party services and Cloud services).

Collaboration

There is a significant overlap in the digital preservation issues being faced by all organisations and across all sectors so it makes sense to pool expertise and experience. There are compelling reasons and, in some cases, political pressure, to engage in greater collaboration within and between organisations in order effectively to confront and overcome the challenges of digital preservation.

Most organisations readily acknowledge the benefits of increased collaboration but also indicate the potential difficulties that can arise in the form of differing agendas, timescales, or funding mechanisms. None the less, it is often possible to collaborate in specific areas or with different levels of intensity that moderate these potential difficulties. Some of the most high-profile and successful initiatives in digital preservation of recent times have been collaborative in nature (see Collaboration).

Organisational change

The modern digital world is a place of both rapid technological and organisational changes. Organisations re-organise internally, merge, or cease to operate with increasing frequency. Digital preservation is a long-term activity and the likelihood of it being affected by organisational change increases over time. This may affect a repository not only through changes to its parent organisation, but through changes to its major depositors and users, suppliers, or collaborators. Organisational change is therefore a major risk to be managed (see Risk and change management).

Organisational structures

The nature of the technology and dependencies in the preservation of digital materials are such that there are implications for organisational structures. Many of the activities converge, for example decisions about acquisition and preservation should sensibly be made at the same time. Organisational structures will need to cross boundaries in order to draw on the full range of skills and expertise required for digital materials. Assigning responsibility for preservation of digital materials acquired and/or created by an organisation will inevitably require involvement with personnel from different parts of the organisation working together. This can potentially present difficulties unless underpinned by a strong corporate vision which can be communicated to staff (see Collaboration, Advocacy, and Staff training and development).

Roles and responsibilities

There are some existing repositories which undertake responsibility for specific subject areas or specific formats. In the UK, for example, the UK Data Service undertakes responsibility for selected social science research data, while the British Library's National Sound Archive assumes responsibility for its collection of sound recordings. Each repository will need to consider its own collection policy and the broader landscape of collecting institutions and remits within which it sits.

The digital environment demands engagement with a large group of stakeholders. The lifecycle approach to digital preservation advocated in the Handbook has significant implications for the way organisations responsible for long-term preservation need to interact and collaborate with creators, publishers and other intermediaries, and each other.

Creators of digital materials need to be able to understand the implications of their actions in terms of the medium to long-term viability of the digital material they create. Whether it be a record created during the day-to-day business of the department, a digital copy of analogue collection material, or a "born digital" resource, guidance and support as well as an appropriate technical and organisational infrastructure will assist in facilitating greatly improved prospects for efficient management and preservation (see Creating digital materials).

Selection

The enormous quantity of information being produced digitally, its variable quality, and the resource constraints on those taking responsibility to preserve long-term access, makes selectivity inevitable if the objective is to preserve ongoing access.

In the digital environment non-selection for preservation may almost certainly mean loss of the item, even if it is subsequently considered to be worthwhile.

In cases where there may be multiple versions, decisions must be made in selecting which version is the best one for preservation, or whether more than one should be selected. Sampling dynamic resources as opposed to attempting to save each change, may be the only practical option but may have severe repercussions if the sampling is not undertaken within a well-defined framework and with due regard to the anticipated contemporary and future needs of the users.

Some consideration also needs to be given in the selection to the level of redundancy needed to ensure digital preservation. There needs to be a clear understanding of who will undertake that responsibility and for what period of time. Otherwise, even if several copies are stored in various repositories, all of those repositories might, for a variety of reasons, cease maintenance of the digital object at some point (see also Acquisition and appraisal).

Balancing security and access

There has always been a strong link between preservation and access. Repositories need to ensure that their digital materials are safe and secure, but most also provide access to a variety of users. Access by real users can provide a valuable steer to the design of preservation facilities, helping to avoid unnecessary actions but also validating and introducing a feedback cycle.

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 other cases there may be legal or regulatory obligations on the repository affecting access. There can be tensions between these two roles and a need to strike a balance between security and ease of access (see Access, and Information security).

Legal compliance

Legal issues are not simple in digital preservation. Multiple copies and derivative versions often exist of digital materials, and there may be associated software and metadata with them from different sources. Digital content is generated by a wider group of creators and incorporates more diverse formats and intellectual property rights (IPR) than applies in the analogue world. The law also often lags behind technological change and digital preservation needs. Some of the key legal issues that affect repositories in collecting, preserving, and providing access to digital materials are:

- Any legal requirements in terms of management, preservation, and access placed upon the repository and its parent organisation, by donors and funders via contracts and agreements or via legislation by Government (e.g. accessibility, availability, information security, retention, audit and compliance, Public Records, Legal Deposit, etc.);
- Those legal obligations relating to third party rights in, or over, the digital materials held by the repository (e.g. copyright, data protection); and
- The legal elements of any relationship between a repository and any third-party provider or providers (e.g. terms of service contracts and service level agreements).

For further guidance and resources to help address these issues and manage associated risks, see Legal compliance and Procurement and third party services sections respectively of the Handbook.

Resourcing Issues

Budgets and costs

The cost of digital preservation cannot be easily isolated from other organisational expenses, nor should it be. Digital preservation is essentially about preserving access over time and therefore the costs for all parts of the digital life cycle are relevant. In that context even the costs of creating digital materials are integral in so far as they may need to include cost elements which will ultimately facilitate their long-term preservation (see Creating digital materials).

The ability to employ and develop staff with appropriate skills is made more difficult by the speed of technological change and the range of skills needed. It is also limited by resource constraints on organisations which may well need to manage growing traditional collections and digital collections without additional resources.

Nonetheless the exercise of calculating costs, however complex, is a valuable and necessary task to establish cost-effective practises and a reliable business model. The cost of the labour required for digital preservation will be the most significant by far and includes not only dedicated experts but varying proportions of effort from many staff such as administration, management, IT support, legal advisers etc.

Other major issues to impact costs include organisational mission and goals, including the type and size of collections, the level of preservation committed to, the quantity and level of access required, and time frame proposed for action. These are discussed in detail in the section on Business cases, benefits, costs, and impact.

The relationship of costs and institutional strategies and activities such as Collaboration, Procurement and third party services, Legal compliance, Staff training and development, or Standards and best practice are also discussed in the relevant sections of the Handbook.

Staffing and skills

Digital preservation involves a range of skills and organisational roles. Typically digital preservation draws on a range of skills which are not normally found in combination. That means larger organisations will likely need to assemble multi-disciplinary teams while in smaller organisations it will be necessary to rely on a distributed team or sources of support.

There are three main issues to consider with respect to staffing and skills:

- Firstly, although there have been considerable improvements in recent years, digital preservation teaching often lags behind current best practice or is wholly theoretical within relevant information management programmes for new entrants into the profession. So individuals with practical skills and experience are in high demand and staff can be hard to recruit.
- Secondly, job descriptions can be hard to script, especially when agencies are effectively starting from scratch with a new role. To this end a number of research projects have attempted to describe generic skills needed for digital preservation, using as a basis the assumption that different skills are required at different levels of an organisation. Tools like the DigCurv Skills framework allied to the Digital Preservation Coalition's Vacancies section can be very useful when describing new roles. Larger organisations with multi-disciplinary teams may be able to recruit to roles that are 'digital' variants of existing professional categories such as archivist, librarian or records manager, but for most organisations new types of roles must be created.

- Finally, staff working in digital preservation frequently report the need to engage in active career development. Given the expectation that technology and the needs of users develop through time, so the staff involved in meeting these changing requirements will need to find ways to have their skills constantly refreshed, such as through specialist briefings and professional networking (see Staff training and development).

Facilities

Effective digital preservation requires some basic facilities or infrastructure, typically technological in nature, on which operational workflows and the processing of digital material can be based. While these may be rudimentary or at least small scale in nature when an organisation takes its first steps in digital preservation, ramping up operations to address large quantities of data will require considerable investment in the facilities required to support it.

Storage

With the typical requirement of replicating preserved data to avoid loss, storage hardware remains amongst the most important digital preservation facilities. Storage technology has changed rapidly over recent decades. Archives widely used media such as CDs or DVDs for long term storage, but the rapid developments in magnetic media have brought fast and reliable storage that has made handheld media redundant. Enterprise storage systems now provide large storage volumes at reasonable cost. While they have finite lifespans, typically of around 4-8 years, they are easy to monitor and then replace when they reach end of life (see Storage).

Organisations may also wish to consider cloud services to "rent" preservation infrastructure. The flexibility of the cloud allows relatively rapid and low-cost testing and piloting. Cloud services can provide easy, automated replication to multiple locations and access to professionally managed digital storage and integrity checking. Repositories can add access to dedicated tools, procedures, workflow and service agreements, providing a digital repository system tailored for digital preservation requirements via specialist vendors (see Cloud services).

Digital repository systems

Many of the core requirements for preserving digital materials are provided in an automated fashion by dedicated digital preservation systems, or trusted digital repositories. A repository application will uniquely identify each digital object placed within it. It will manage the storage of that object, identify its characteristics and help a repository manager to plan its preservation. It will also facilitate access to the object. While basic preservation can be provided on an ad hoc basis at a small scale, a dedicated repository application is essential to managing digital materials effectively over time. The OAIS model provides a high level model for the functions required by a repository (see Audit and certification for more information on certification of trusted digital repositories and Tools for repository systems and components).

High performance computing

Increasing volumes of data require not only more storage but also greater computational power. Characterising and assessing the technical characteristics of data, indexing data to enable search and access, integrity checking and a host of other tasks require considerable computational performance. Those dealing with these big data, be it research data or web archives have typically looked to high performance computing, and technologies such as Apache Hadoop running on clusters of commodity hardware to meet this need.

Digital preservation laboratory

A number of larger organisations have developed lab environments within which an array of old and new technology can be applied for the stabilisation or ripping of data from obsolete media, and has been championed by organisations working with personal digital collections. Specialist drives for reading magnetic media, robots for processing large numbers of optical disks and write blockers for allowing access to hard drives without changing the bits in the process, are just some of the equipment that could be useful here. Media recovery companies offer an alternative approach that may be preferable in high volume cases, albeit with less control of the process and the need to move media offsite (see Digital forensics).

Resources

How Toy Story 2 Almost Got Deleted: Stories From Pixar Animation: ENTV

https://www.youtube.com/watch?v=8dhp_20j0Ys

Entertaining and informative story of how 'Toy Story 2' was almost deleted from Pixar Animation's computers during the making of the film and how the film was saved by one mom's home computer (2 mins 26 secs)

References

Rosenthal, D., 2014. *Talk "Costs: Why Do We Care?", DSHR's Blog*, Tuesday November 18 2014. Available: http://blog.dshr.org/2014/11/talk-costs-why-do-we-care.html

Digital Preservation **Handbook**

Getting Started

Illustrations by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Gold sponsor

Silver sponsors

Bronze sponsors

Reusing this information

Contents

Getting Started

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Introduction

This section is for you if you have yet to start digital preservation or have just begun to do so. It provides a rapid introduction to a number of approaches that will support you in getting started, moving towards using other sections of the Handbook, and in building confidence and skills.

The section developed out of the "Getting Started in Digital Preservation" workshops run by the Digital Preservation Coalition. It supports 'learning by doing' and assumes a minimal level of prior knowledge. As you progress you will benefit from dipping into the resources and case studies, other topics and sections in the Handbook, and the Glossary for any unfamiliar terms.

Digital preservation can seem like a daunting prospect. It can help to map out the skills and resources you have and the materials you want to preserve. That way you start with what you know rather than what you don't. The first step in digital preservation is almost always to undertake a rapid assessment. This will have two or three components:

- knowing the practical capacity of your organisation;
- understanding the organisation's goals and missions; and
- knowing a little about the digital materials in question.

Get to Know Your Organisation and Your Data

Creating a Digital Asset Register

As part of a rapid assessment: it is vital to understand the nature and extent of your digital collections. A digital asset register will be incredibly useful for assessing the extent and significance of the collection, identifying priorities and planning digital preservation actions. A high-level assessment of the collection will help with more detailed mapping later: a comprehensive and detailed audit could be time consuming. So the advice in early stages is to keep the asset register simple. Ask the following questions:

- What is the subject of the collection?

- Where does it come from and what is its function?
- Where is it stored and what kinds of media are used?
- Why is it being retained?
- Who is responsible for it; who are the users; who are the subjects of the data?
- How is the data accessed?
- How is the data likely to change and grow in the near future?

Assessing Your Organisation's Readiness

Organisational maturity is another factor to consider. The National Digital Stewardship Alliance (NDSA, 2013) in the USA has recommended a simple 4 level model to help organisations understand and improve their technical capacity in digital preservation. The four levels are:

- Level 1 - protect your data
- Level 2 - know your data
- Level 3 - monitor your data
- Level 4 - repair your data

These 'Levels of Preservation' are intended to be progressive, and are used to measure maturity against four components: storage, file fixity, information security, metadata, and file formats. An organisation's capacity to undertake digital preservation is indicated by its maturity level across these five components. More comprehensive maturity models are available, such as the Digital Preservation Capability Maturity Model (Dollar and Ashley, 2014), if a more well-rounded exploration of organisational maturity is required.

First Steps to Securing Your Data

This section provides an overview of initial actions to secure your data once you have assessed your organisation's readiness and compiled basic information about your data. The following steps are essential to ensuring a minimum level of preservation when a new collection of digital material is received. This is typically referred to as bit preservation. Quite literally, preserving the streams of binary digits, or bits, that make up your digital files (without preserving the means to decode the bits into meaningful information).

Prompt check in on receipt

When a new collection of digital material is received from a supplier it is essential to ensure that what has been received is what is expected. Depending on the source of the material, it may be possible to request new copies of missing or poor quality files. These checks are made and any replacement requests submitted, the greater the likelihood of successful resolution.

Key tasks include:

- Scan for viruses and malware to make sure there are no unwanted surprises in the collection. Perhaps keep the collection 'in quarantine' until you have checked it.
- Check all expected files are present. If the material is accompanied by a manifest, check the files against it

- Open a random selection of files to verify their integrity and/or expected quality levels
- Promptly request replacements for any damaged or missing files, where possible

Create a Verifiable File List

In order to check over time that your digital files are being preserved, it is first necessary to record exactly what files are in your possession. It is, therefore, important to create a verifiable list of files in each collection. These lists should likely contain information such as file names, locations and sizes, format types and checksums. A checksum is a short alphanumeric string that represents the contents of a file which acts as a 'digital fingerprint' allowing comparison over time. Once the list has been created, it is a simple process to verify that all files are present and undamaged, at any point in the future (see fixity checking below).

Various software tools can be used to automatically generate this data; these are commonly referred to as characterisation tools. For example, you may wish to use The National Archives PRONOM (a register of file formats and their behaviours) and DROID (a tool that uses PRONOM to analyse the files on a system). Having a list of the file formats, versions and quantities in your collections will help you make a case to senior management for the support and resources that will be needed to do the job properly and sustainably. This information can also be used to update and enrich your digital asset register. The range of formats in use should be consolidated to minimize duplication and eliminate problem formats. This process is known as normalization.

Key tasks:

- Generate a verifiable file list
- Update digital asset register

Stabilise your files: make copies

No matter how good your digital storage, your digital material will always be at risk of damage, decay or accidental deletion. Making more than one copy of your digital materials and utilising more than one type of storage solution mitigates a variety of digital preservation risks.

Key tasks include:

- Keep (at least) one copy easily accessible on non-removable disk. You will need to regularly revisit your material to ensure its fixity, so keeping it accessible will make this easier
- Make (at least) one additional copy, if necessary on a less accessible, but cheaper storage medium such as tape
- Keep one copy in a different geographical location to the others to mitigate against disaster

Revisit and inspect: Fixity checking

By revisiting your digital materials on a regular basis (e.g. every 6 months) you can ensure that no damage or accidental loss has occurred. If it has, you can recover problematic files from the copies or backups you have made previously. Future fixity checks will generate new digital fingerprints (or checksums) for the files in your collections. If they do not match the ones originally created, bit loss or damage has occurred.

Key tasks include:

- Revisit your collection on a frequent basis, recalculate the checksums, identify files that have become damaged

- Retrieve copies of damaged files and repair as necessary
- Perform test recoveries of data backed up by third party services, to ensure backups are being performed as agreed

Document your processes

From the outset of creating a digital collection, it is important to document as much as possible about a collection's assets, the tools and workflows. This documentation is an important component of technical and descriptive Metadata. It is necessary to retain this information for the purposes of longevity. As, with any project, staff retention can be an issue. If staff leave they often take essential knowledge and skillsets.

Where next?

Having taken the first steps in digital preservation, where do you go next? This will obviously depend on your own requirements and priorities, but this table provides a number of suggestions and other sections of the Handbook will help you move forward with them:

Next steps
Develop advocacy and outreach, an understanding of risk, the business case, costs, benefits and impact
Establish an organisational preservation strategy and policies. As well as ensuring a consistent approach to preservation it can be a useful tool to achieve buy in across an organisation and in particular with senior management
Establish a digital repository. Technical solutions and tools either on local IT infrastructure or offered as a cloud service will help you understand, manage and preserve your digital material for the long term
Establish your long-term storage, preservation planning and action
Revisit and expand your collection audits: • Characterise priority collections in more detail • Periodically update collection audits as required
Establish a digital preservation working party. Effective digital preservation often requires buy in across many departments within an organisation. A representative working party can be vital in making coordinated steps forward
Build the necessary staff training and development and skill sets
Establish a professional network and collaborations. Join a digital preservation membership organisation such as the Digital Preservation Coalition

Keep up to date with new developments:

- Email lists for digital preservation include the digital preservation announcement list on JiscMail, and the USA-focussed digipres list
- A weekly DP News blog selects recent tweets and news links on digital preservation
- Journals with a digital preservation focus include: International Journal of Digital Curation, and D-Lib
- Events with a digital preservation focus include iPRES, and PASIG
- The Digital Preservation Coalition organises briefing days on particular digital preservation topics

Resources

A Preservation Primer

http://knconsultants.org/a-preservation-primer/

This clear practical short primer on preservation for beginners was written by staff at Portico. It summarizes the issues and outlines various short and long-term preservation options that an organization might take to begin planning for long-term digital preservation of its content, beginning with near-term protection and concluding with full preservation and long-term protection. (83 pages).

Don't Panic:The Archivist's Guide to Digital Preservation

http://www.wyjs.org.uk/documents/archives/dont_panic_digital_preservation_first_steps_guide.pdf

A practical and concise guide produced in 2011 by Stefanie Davidson at the West Yorkshire Archive Service. It is intended to act as a signpost to assist in taking the first steps in understanding some of the issues involved rather than a comprehensive guide and an introduction to the topic to help you find your feet. (8 pages).

Putting Parsimonious Preservation into Practice

http://www.nationalarchives.gov.uk/documents/information-management/parsimonious-preservation-in-practice.pdf

The principle of Parsimonious Preservation was originally developed in 2009 at The National Archives in the UK as an approach for small or medium sized institutions to permit them to begin work on digital preservation but is also practical for large scale institutions. It now underpins advice and guidance given to the UK archive sector on digital preservation. (11 pages).

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, and Preserving digital Objects With Restricted Resources project (POWRR). 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

DPC Getting Started in Digital Preservation Workshops

http://www.dpconline.org/events

The DPC Getting Started in Digital Preservation workshops are events designed to raise awareness of digital preservation issues, increase involvement with digital preservation activities and sign-post the support and resources available to help you on your way. They provide an introduction to digital preservation, build an understanding of the risks to digital materials, include practical sessions to help you apply digital preservation planning and tools, and feature speakers sharing their own experience of putting digital preservation into practice. You can find details of forthcoming workshops and the programmes and speaker presentations at previous workshops on the DPC events page.

Digital Preservation Management: Implementing Short-Term Strategies for Long-Term Solutions

http://www.dpworkshop.org/

An excellent free online tutorial that introduces you to the basic tenets of digital preservation. It is particularly geared toward librarians, archivists, curators, managers, and technical specialists. It includes definitions, key concepts, practical advice, exercises, and up-to-date references. The tutorial is available in English, French, and Italian.

Canadian Heritage Information Network (CHIN) Digital Preservation Toolkit

http://www.rcip-chin.gc.ca/carrefour-du-savoir-knowledge-exchange/outils_preservation_numerique-digital_preservation_toolkit-eng.jsp

CHIN has released a suite of documents to identify digital material found in museums, the potential risk and impact of lost material, and how to get started in the development of Preservation Policies, Plans and Procedures. The toolkit includes a Digital Preservation Inventory Template, Digital Preservation Policy Framework Development Guideline, Decision Trees, and a Digital Preservation Plan Framework.

Digital Preservation 101, or, How to Keep Bits for Centuries

http://scholar.harvard.edu/jcs/presentations/dhttp://handbook.dpconline.org/administrator/index.php?option=com_content&view=article&layout=edit&id=86igital-preservation-101-or-how-keep-bits-centuries

This 2015 presentation by Julie Swierczek Digital Asset Manager and Digital Archivist at Harvard University Art Museums is a good advocacy for and explanation of, digital preservation to othernon-specialist institutional colleagues including "why archivists cry themselves to sleep at night when the general public conflates archives with backup copies of data" (142 slides but many are images with good slide notes that make this easily understandable).

The National Archives Digital Continuity Guidance

http://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/step-by-step-guidance/step-2/

This guidance takes you through the process of creating an information asset register, and includes a template in Excel spreadsheet format. The register can be useful for Records Managers/Information Managers as a model which they can demonstrate aligns with business risk management.

Risk Management for Digital Preservation

https://vimeo.com/171082277

From a series of video covering topics from the 'Getting Started in Digital Preservation' roadshows, this video provides a brief introduction to the use of risk management for Digital Preservation.

Case studies

Bishopsgate library case study

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

A collections audit and business case focused on taking the first steps of digital preservation at the Bishopsgate Institute Library. (28 pages).

Starting Small: Practical First Steps in Digital Preservation

http://www.slideshare.net/hakbailey/starting-small-practical-first-steps-in-digital-preservation-13385434

One example of how digital preservation principles can be added to the collections management activities of a small institution (Dartmouth College USA from 2010–2012), without needing a lot of additional resources. (26 slides).

DPC case note: West Yorkshire Archive Service accepts a digital collection

http://www.dpconline.org/component/docman/doc_download/511-casenotemlawyas.pdf

In this Jisc-funded case study staff from West Yorkshire Archives Service report on their experience in taking their first large digital archive. This made them confront new problems and new ways of working, they conclude that "If we try we may fail; if we don't try we will certainly fail". October 2010 (4 pages).

DPC case note: Glasgow Museum takes first steps in turning an oral history headache into an opportunity

http://www.dpconline.org/component/docman/doc_download/502-casenoteglasgowmuseums.pdf

This Jisc-funded case study examines how Glasgow Museums' took some simple steps in addressing digital preservation and created short and long term opportunities. Activities such as creating an inventory, assessing significance and promoting access provide the basis for building confidence to manage the wider challenges, and can bring early rewards if properly embedded within the mission of an organization. September 2010 (4 pages).

Digital Preservation Planning Case Study

http://www.dpconline.org/component/docman/doc_download/863-2013-may-getting-started-london-planning-case-study-ed-fay

A set of DPC Getting Started in Digital Preservation workshop presentation slides by Ed Fay from May 2013. An excellent concise overview of planning for digital preservation and how to approach it . (20 slides).

References

Dollar, C.M. and Ashley, L.J., 2014. *Assessing Digital Preservation Capability Using a Maturity Model Process Improvement Approach.* Available: http://static1.squarespace.com/static/52ebbb45e4b06f07f8bb62bd/t/53559340e4b058b6b2212d98/1398117184845/DPCMM+White+Paper_Revised+April+2014.pdf

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

Digital Preservation **Handbook**

Institutional Strategies

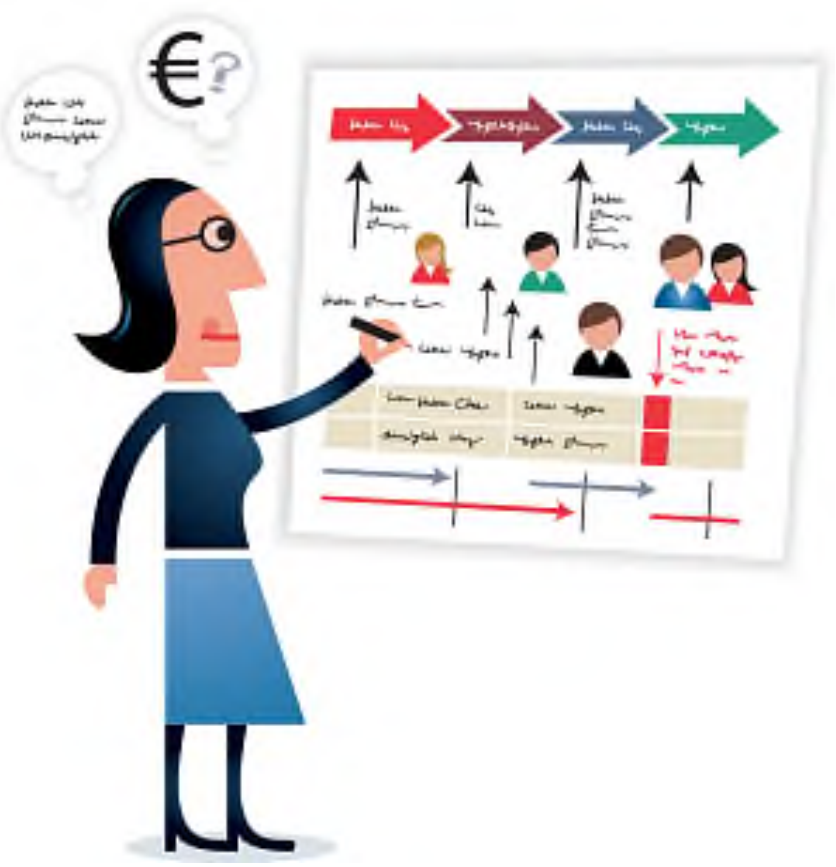


Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Who is it for?

Both senior administrators (DigCurV Executive Lens) and operational managers (DigCurV Manager Lens) within institutions. Also existing or potential third-party service providers.

Assumed Level of Knowledge

Intermediate (basic understanding of the issues, some practical experience).

Purpose

- To form the basis for further development of policies and strategies appropriate to individual institutions.
- To provide existing examples of good practice which might serve as models.
- This section outlines a number of strategies which have been used successfully by institutions in developing approaches to digital preservation. Each sub-section discusses the approach, its potential advantages and disadvantages, and then provides exemplars of the approach together with further reading on the topic. Strategies such as these will form a core component of corporate policy development to address digital preservation. Sound policy development combined with effective working practices and procedures (see Organisational activities) has been essential to effective digital preservation programmes.

Reusing this information

Contents

Institutional Policies and Strategies

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Introduction

The aim of this section is to help institutions understand, develop and implement digital preservation policies and strategies. These will help an organisation to set digital preservation goals, priorities and mechanisms that will also support the acquisition, life cycle management and dissemination of digital materials.

Policy and strategy are terms that are often used interchangeably or in different hierarchical sequence in different institutions. For consistency, the Handbook defines 'policy' as the highest level document and 'strategy' as the documents and procedures that support the implementation of the policy. In principle the development of policy precedes the development of strategy. In turn strategy may be developed or revised/reviewed on a regular basis, whereas policy may have a longer review cycle. Thus a policy serves the organisational need whilst individual strategies may serve different business units or divisions.

Policy and strategy documents provide a foundation upon which all activities around management of digital materials can be based. Policy and strategy documents that are well formed and consultative provide for high levels of both consensus and compliance in the day-to-day activities of managing digital materials. In turn, this provides for certainty that digital materials are being managed appropriately and to best effect. Policy documents also form the basis for cost planning and for funding applications. Strategy can be used as a flexible means to both adapt to changing situations and to demonstrate that learning that has been applied.

Within any institution there will be a range of stakeholders who have a stake in the life cycle management of digital materials. They may contribute to the management of those materials, they may create or manage metadata associated with those materials, or they may have management responsibility for collections. The end-users are also key stakeholders as their needs determine what is important for preservation. The views of stakeholders and their roles in relation to the management of digital materials must be considered at both the policy and strategy level.

You may find it useful to apply the iterative four-step management method of Plan–Do–Check – Adjust as a model for continuous improvement and effective development, implementation and revision of policies and strategies.

Digital preservation policy as part of the wider organisational context

If you are embarking upon, or thinking of embarking upon the creation of a digital preservation policy for your organisation, then it is necessary to start by investigating the context in which the policy will exist.

It is likely that a broad range of policy documents will already exist across your organisation covering a wide variety of issues such as staffing, information technology, risk assessment, and finance. There mayalso be a number of policies relating to more specific issues of records and collections management that will be relevant to digital preservation activities. It is essential to consider both the content and established style and structure of all relevant policies within the organisation as well as the how digital preservation policy will fit within the wider landscape. No single policy or strategy document can stand alone so to achieve support for and successful implementation of a digital preservation policy it is essential to embed it in the broader policy context.

Important considerations in developing policy and strategy

An important aspect of policy development is consideration of the specific needs of your organisation and its key drivers. Alignment with organisational business drivers ensures that strategy and its implementation are also aligned with business need. Policy and strategy documents should make explicit links with and between relevant and existing policies and strategies and build on existing practice. Collaboration, sharing and consultation with stakeholders are essential processes in the development of policy and strategy.

Digital preservation policies are ideally technology neutral, i.e. not dependent upon any one technology platform or system. However in reality this may be unachievable. In such cases they should be focused on principles, aims and objectives that the requisite technology can support.

In order to develop clear, coherent and robust documentation and processes it is essential to adhere to a set methodology and to establish a plan for review so that the policy and strategy remains relevant and current.

1. **Establish purpose**. The first step is to establish the main purpose of the digital preservation policy, its scope and key aims. These will keep the process of policy development focused and its content coherent. Thought should be given at this stage to how the document will used, both as a tool for advocacy and to help guide the creation and implementation of strategy.

2. **Research**. As expressed above, it is essential to understand the organisational context in which the policy will exist. Time should be spent investigating existing policy, understanding the organisation's business drivers and the needs of key stakeholder groups. This phase will also incorporate research into best practice for digital preservation policy and strategies, examining the tools and resources available as well as reading policies and strategies from other organisations. Many resources are available with suggestions of what to include in your digital preservation policy and strategy (see Resources).

3. I**dentify elements and develop structure**. Based on the research carried out in the previous phase, the main topics and issues to be addressed in the policy and strategy should be selected. Developing a clear structure for the documents is essential to ensure the

documents are useful in practice and to facilitate easy updates and review. The structure should reflect any standards or existing best practice for policy and strategy documents within the organisation.

4. **Develop content**. Policy content should be high-level and set broad aims and objectives. It should avoid identifying specifics such as details of particular technology solutions, although it may contain reference to commitments established on an organisational level. Information on practical application of the policy will be defined by relevant strategy documents. Content may also be aspirational in relation to aims and objectives but care must be taken not to set unobtainable goals. Recommendations on how to address specific issues within your policy and strategy are available from numerous sources (see Resources).

5. **Stakeholder review**. It is essential to gain buy-in from various stakeholder groups to ensure your policy and strategy are both fit for purpose and will have support from across the organisation. Presentation of the draft documents to key stakeholder groups is an important part of the drafting process and any feedback provided should be considered carefully. This can also be a key step in advocacy for digital preservation within your organisation, allowing stakeholders to feel engaged with the process and to understand how digital preservation activities relate to their own work. (see Advocacy)

6. **Gain approval**. Most organisations will require that new policy documents are officially ratified by your management board. Make sure to be aware of the process the organisation and any requirements that will need to be fulfilled. Once ratified the policy will carry more weight and as a result will be easier to implement as part of ongoing strategy.

7. **Regular reviews**. Policy and strategy documents should not be static and must be responsive to changes in stakeholder needs, the wider organisational context and updates to best practice. A regular review cycle should be established but may also be triggered by significant changes in any of the areas mentioned above.

8. **Implementation**. Establish an implementation plan to make policy and strategy a reality in terms of day-to-day operations. Remember they are a mean to an end, not an end in itself.

Resources

Digital Preservation Policies Study

http://www.webarchive.org.uk/wayback/archive/20140615022334/http://www.jisc.ac.uk/media/documents/programmes/preservation/jiscpolicy_p1finalreport.pdf

This JISC funded study published in 2008 created a model framework for a digital preservation policy and accompanying implementation clauses based on examination of existing digital preservation policies. Although focussing on the UK Higher and Further Education sectors, the study draws widely on policy and implementations from other sectors and countries.

An additional output was a series of mappings of digital preservation links to other key institutional strategies in UK universities and colleges with the aim of helping institutions and their staff to

develop appropriate digital preservation policies and clauses set in the context of broader institutional strategies. (60 pages).

Digital Preservation Policies: Guidance for archives

http://www.nationalarchives.gov.uk/documents/information-management/digital-preservation-policies-guidance-draft-v4.2.pdf

This guide published by The National Archives in 2011 explains the key characteristics of a digital preservation policy. It discusses why there is a need for a policy and how it supports digital preservation. The primary audience for the guidance is publicly funded archives. (16 pages).

Analysis of Current Digital Preservation Policies: Archives, Libraries and Museums

http://www.digitalpreservation.gov/documents/Analysis%20of%20Current%20Digital%20Preservation%20Policies.pdf?loclr=blogsig

This report published in 2013 by Madeline Sheldon, a Junior Fellow with NDIIPP at the Library of Congress, discusses the current state of digital preservation policy planning within cultural heritage organizations. The collection of new or recently revised digital preservation policies or strategies, published during 2008 and 2013, resulted in a high-level analysis of the contents within those documents. A summary overview of the findings was also made available as a post on The Signal blog. (23 pages).

APARSEN D35.1 Exemplar good governance structures and data policies

http://www.alliancepermanentaccess.org/index.php/consultancy/member-resources/documents-and-downloads/?did=174

This report summarises the level of preparedness for interoperable governance and data policies. It concludes with selected recommendations that should be taken into account when drawing up data policies concerning digital preservation. (2014, 43 pages).

SCAPE Catalogue of Preservation Policy Elements

http://wiki.opf-labs.org/display/SP/Catalogue+of+Preservation+Policy+Elements

The European Project SCAPE (2011 - 2014) was tasked with looking at policy and producing a catalogue of policy elements to assist those writing policy. This wiki gives some information on the background to the policy work and then has pages for each element which the SCAPE project suggested that organisations should consider when writing policy, with a focus of planning and watch activities. There is also the final report of this work made publicly available in February 2014 on the SCAPE website.

Published Preservation Policies

http://wiki.opf-labs.org/display/SP/Published+Preservation+Policies

An extensive web directory prepared by the SCAPE project in 2015 listing digital preservation policies that are publicly available online for libraries, archives, data centers, and miscellaneous institutions.

Case studies

A Digital Preservation Policy for Parliament

http://www.parliament.uk/documents/upload/digitalpreservationpolicy1.0.pdf

The purpose of this Policy published in 2009 is to state and communicate the principles that guide the UK Parliament's activities to secure the preservation of its digital information resources. Further policy documents, procedures, standards, and guidance will be developed in future to address specific aspects of the Strategy. (17 pages).

Hampshire Archives and Local Studies (HALS) Digital Preservation Policy

http://www3.hants.gov.uk/archives/hro-policies/hro-digital-preservation-policy.htm

To address the risk of losing digital materials, HALS has developed a digital preservation policy and strategy. The policy outlines the Record Office's approach to digital preservation, whilst the aim of the strategy is to describe this approach in more detail, including technical specifications where appropriate.

DPC case note: Cabinet papers - policy as a measure of commitment

http://www.dpconline.org/component/docman/doc_download/449-casenotecabinetpapers.pdf

This case note from The National Archives examines the relationship between policy and practice in digital preservation. Grant giving organisations should request copies of applicant's digital preservation policies when funding data creation, as these are an indication of the organisation's commitment to long-term access. The National Archives has digitised a significant volume of the UK's Cabinet Papers, and have a carefully considered policy framework for the long term management of digital resources. May 2010 (3 pages).

DPC case note: Welsh journals online: effective leadership for a common goal

http://www.dpconline.org/component/docman/doc_download/450-casenotewelshjournals.pdf

This Jisc-funded case study examines a complex digitisation project at the National Library of Wales, an example of an organisation where there are many stakeholders and many different skills are required. Nominating a single senior member of staff as the lead officer for digital preservation and allowing them to work across different sections of the institution mitigated the risk of uncertainty around responsibility for preservation actions. June 2010 (3 pages).

British Library Digital preservation strategy 2013-2016

http://www.bl.uk/aboutus/stratpolprog/collectioncare/digitalpreservation/strategy/dpstrategy.html

The British Library's Strategy includes four priorities. Each priority is accompanied by a series of actions. These priorities are aligned with the Library's overall approach to Collection Care and its five principles of sustainable stewardship: to predict, protect, prioritise, preserve, and enable.

Strategic Priority 1: Ensure our digital repository can store and preserve our collections for the long term

Strategic Priority 2: Manage the risks and challenges associated with digital preservation throughout the digital collection content lifecycle
Strategic Priority 3: Embed digital sustainability as an organisational principle for digital library planning and development
Strategic Priority 4: Benefit from collaboration with other national and international institutions on digital preservation initiatives

Further details of each can be found in the full version of the strategy in a linked pdf document (16 pages).

Wellcome Library's Preservation Policy

http://wellcomelibrary.org/what-we-do/library-strategy-and-policy/preservation-policy/

The purpose of the Wellcome Library's Preservation Policy is to provide a comprehensive statement on the preservation and conservation of the Library's collections. It is intended to cover all material in all formats. The policy contains three parts that cover general statements, the management of physical materials and the management of digital materials.(25 pages).

UK Data Archive Preservation Policy

http://data-archive.ac.uk/media/54776/ukda062-dps-preservationpolicy.pdf

This policy published in 2014 outlines the principles which underpin the main activities of the UK Data Archive (the Archive) the active preservation of digital resources for use and re-use within its core user community. From a preservation point of view this policy generally conforms to the OAIS Reference Model, with additions and alterations which are specific to the materials held within the Archive. The Archive has a series of strict requirements for its digital preservation activities. These requirements are laid down in this policy, and the manner in which these requirements can best be achieved in relation to regulatory requirements, archival best practice, information security and available funds is also detailed below. Consequently, the Archive's preservation policy is based upon open and available file formats, data migration and media refreshment. (16 pages)

Digital Preservation Strategies for a Small Private College

http://files.archivists.org/pubs/CampusCaseStudies/CASE-16-MegMiner-Final.pdf

The POWRR Project (2011 – 2014) investigated, evaluated, and recommended scalable digital preservation solutions for libraries with smaller amounts of data and/or fewer resources. Well established "best practices" in digital preservation (DP) do little to address day-to-day realities in repositories that cannot dedicate funds or staff to DP workflows. Meg Miner, Illinois Weslyan University, discusses what can be done to ensure good stewardship for born digital and digitized institutional records before a complete preservation system is in place. 2015 (13 pages).

University of Edinburgh - Developing a Digital Preservation Policy

http://www.dpconline.org/component/docman/doc_download/1321-making-progress-hsbc-nov-2014-lee

A great presentation and case study by Kirsty Lee at the University of Edinburgh to the DPC Making Progress in Digital Preservation workshop in October 2014 explaining the methodology that she is using to build a digital preservation policy at Edinburgh. (14 pages)

Collaboration

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Introduction

There are compelling reasons and, in some cases, political pressure, to engage in greater collaboration within and between organisations in order effectively to confront and overcome the challenges of digital preservation. The range of skills required to do this demands flexibility within organisational structures to facilitate working in multi-disciplinary teams. There is a significant overlap in the digital preservation issues being faced by all organisations and across all sectors so it makes sense to pool expertise and experience. Communication with key stakeholders, using terms and language understood by them (see Advocacy) will play a major part in building and maintaining collaborations.

Internal collaboration

The usual assumption is that collaboration is external. However, most libraries and archives will be managing a combination of paper-based and digital resources for the foreseeable future and will need to structure their organisation to manage the disparate needs of the two. The blurring of boundaries and the lifecycle changes which digital technology produces means that sections and departments which are structurally distinct, will now need to co-operate in order to integrate the preservation and management of digital materials with other materials.

Such co-operation and joined-up working may well prove impossible unless there are mechanisms put in place to facilitate it and there is clear executive buy-in and sponsorship to promote action. At the strategic level, a cross disciplinary committee or project team charged with developing and overseeing objectives is one means of ensuring that all relevant sections can be brought together. At the operational level, consideration will need to be given to defining what specific tasks are required and where those responsibilities logically lie. Setting up of working groups to investigate specific issues is one means of blending the range of skills required. Good communication and advocacy to other stakeholders will also be important (see Advocacy).

Whilst the need for a policy or strategy relating to digital preservation may be well established within the repository team, the driver for internal collaboration is typically in response to a specific challenge faced by an organisation.

Advantages

- Makes good use of available skills and expertise and makes the case that digital preservation is an institutional issue and not one owned exclusively by the repository.
- Promotes cross-team working by improving understanding of shared objectives and who needs to contribute.
- The sooner digital preservation becomes part of the daily work of an organisation and its employees, the better it is for their transition to and readiness for a more digital world.
- Recognises the diversity of skills required for the digital environment in general and digital preservation in particular.
- Is more likely to be focused and aligned with institutional objectives and priorities.
- Maintains a high profile for the work.

Disadvantages

- May be frustrating and time consuming in the short term.
- Communication may be difficult initially - for example there are some issues surrounding terminology with the term 'archives' meaning different things to archivists and IT colleagues
- Senior management may be unwilling to risk perceived lack of control.
- Staff may feel uncomfortable with new ways of working.
- Organisational structures may not be sufficiently flexible to facilitate effective collaboration between different sections.

External collaboration

There may be a number of drivers for external collaboration. There is the simple desire for lone specialists to work with other professional colleagues and seek external validation of their ideas or direction of travel. At the other end of the scale is the response to external funding opportunities, with funders now placing greater emphasis on collaboration. Some examples of types of external collaboration in the digital preservation sector are included below:

- **Collaboration around a specific problem to make progress easier and more affordable**. The Digital Preservation Coalition itself is an example of this in the UK. Members are encouraged to engage and collaborate on a number of different digital preservation related issues both at a high level and on specific topics. Another example is the Section for Archives and Technology of the Archives and Records Association, which brings together members of the professional body to look at specific aspects of the work and to share current practice.
- **Collaboration around a standard**. An example of this would be the call in 2015 to work together around the revision of the OAIS reference model. In an initiative coordinated by the DPC, practitioners working in the field were encouraged to engage and feed into a shared response. (See http://wiki.dpconline.org/index.php?title=OAIS_Community)
- **Collaboration around a specific piece of software or system**. An example of this would be the user groups that evolve around digital preservation software solutions, both commercial and open source. When exploring a piece of software for the first time there is huge value in being able to share experiences and learn from others.

- **Collaboration within a specific geographical area**. There are many examples of organisations collaborating based on their geographical proximity and the ease of working together that this offers. One example of this is the Digital Preservation Group within Archives & Records Council Wales (ARCW) (see Case studies).

Advantages

- Organisational commitment and authority.
- Formal agreements offer a clear allocation of responsibilities between partners.
- Enhanced understanding of complex issues.
- Greater practical benefit from pooled resources and expertise.
- Enhanced reputation through successful delivery of a project or being able to manage digital preservation.
- Improved prospects for future mutually beneficial collaboration.

Disadvantages

- Difficulty of establishing unambiguous agreements able to be accepted by all parties.
- Time taken to establish teams or a collaborative framework.
- Difficulties of communicating across different professional and organisational frameworks.
- Potential bureaucratic barriers.

External collaboration can work on an informal or a formal basis and colleagues across the sector have always shared experiences, with informal collaboration often forming part of an individual's continuing professional development. Larger more complex collaborations are more likely to have a formal partnership agreement that can be useful to define the scope and boundaries of the working relationship and attribute specific responsibilities.

Resources

Benefits from Research Data Management in Universities for Industry and Not-for-Profit Research Partners

http://opus.bath.ac.uk/32509/

Applies a stakeholder mapping using the KRDS Benefits Framework to examine the data management benefits associated with Faculty-Industry and Faculty-Not-for-Profit research collaborations with the University of Bath. It presents a summary list of benefits to different stakeholders that can arise from research data management and data preservation in these collaborations.

Aligning National Approaches to Digital Preservation Conference Proceedings 2012

http://educopia.org/publications/anadp

This publication contains a collection of peer-reviewed essays that were developed by conference panels and attendees. It aims to establish a set of starting points for building a greater alignment across digital preservation initiatives and highlights the need for strategic international collaborations to support the preservation of our collective cultural memory (342 pages).

North West Region Digital Preservation Group

https://nwrdpg.wordpress.com/

The North West Region Digital Preservation Group is an example of an informal geographical collaboration involving local authority, academic and specialist archivists. Outcomes include guidelines for depositors, a workbook for archivists and pilot studies on web archiving and email archives.

Case studies

DPC case note: Welsh journals online: effective leadership for a common goal

http://www.dpconline.org/component/docman/doc_download/450-casenotewelshjournals.pdf

This Jisc-funded case study examines a complex digitisation project at the National Library of Wales, an example of an organisation where there are many stakeholders and many different skills are required. Nominating a single senior member of staff as the lead officer for digital preservation and allowing them to work across different sections of the institution mitigated the risk of uncertainty around responsibility for preservation actions. June 2010 (3 pages).

DPC case note: Freeze Frame preservation partnerships

http://www.dpconline.org/component/docman/doc_download/434-casenotefreezeframe.pdf

This case note examines the relationship between the relatively short lived Freeze Frame project at the Scott Polar Research Institute and the institutional repository which offered to provide long term preservation services to ensure ongoing access at the end of the project. The study shows that small organisations don't necessarily need to establish a sophisticated preservation infrastructure when they embark on digitisation. Partnership can bring unexpected benefits to both parties, but needs to be thoughtfully managed and documented. April 2010 (4 pages).

Community Action via UK LOCKSS Alliance

http://www.slideshare.net/edinadocumentationofficer/ukla-dpc-final

Presentation given by Adam Rusbridge at the Digital Preservation Coalition on Getting Started in Digital Preservation, 28 February 2011. It discusses the role of the UK LOCKSS Alliance and collaboration in e-journal preservation.

Archives & Records Council Wales Digital Preservation Working Group

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

This National Archives 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. January 2015 (10 pages).

A collaborative infrastructure for permanent access to digital heritage in The Netherlands

http://www.ncdd.nl/wp-content/uploads/2016/03/Network_Digital_Heritage_Netherlands.pdf

In 2014 the Network Digital Heritage (NDE) was set up in 2014 by a group of national organizations in the Netherlands. The network presented a strategy for the development of a national, cross-domain infrastructure of digital heritage facilities. One of the programmes focusses on digital preservation (Sustainable digital heritage). The aim of this programme is to work on the cross-sector sharing, utilisation, and scaling up of facilities for sustainable preservation and access, while devoting attention to cost management and the division of duties. This programme is carried out by the NCDD, the National Coalition for Digital Preservation. (3 pages).

The SPRUCE project

http://wiki.opf-labs.org/display/SPR/Home

The Sustainable PReservation Using Community Engagement (SPRUCE) project (2011-2013) sought to inspire, guide, support and enable HE, FE and cultural institutions to address digital preservation gaps and to use the knowledge gathered from this activity to articulate a compelling business case for digital preservation. This multi-institutional collaboration brought archivists and technology experts together through mashup events and a hackathon. Two key outputs from the project were the Business Case Toolkit (http://wiki.dpconline.org/index.php?title=Digital_Preservation_Business_Case_Toolkit) and COPTR (Community Owned digital Preservation Tool Registry) (see http://coptr.digipres.org/Main_Page).

Filling the Digital Preservation Gap Case Study

http://digital-archiving.blogspot.co.uk/2015/12/research-data-spring-case-study-for.html

A collaboration between the Universities of Hull and York. The aim of the project was to address a perceived gap in existing research data management infrastructures around the active preservation of the data. Both Hull and York had existing digital repositories and sufficient storage provision but were lacking systems and workflows for addressing the active preservation of data.

Advocacy

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Introduction

Digital preservation relies on a wide range of skills and services, so digital preservation managers need to coordinate a diverse set of skills, policies, tools and services from disparate sources. For some organisations digital preservation is entirely new and the relevant resources will need to be assembled for the first time. Even established programmes will face new challenges and therefore the range of tools and services required may constantly change. Hence the ability to communicate with other staff, departments, and organisations has emerged as a key skill for successful digital preservation managers.

Because technology and staff continue to change, communication and advocacy must be an ongoing rather than a one-off activity.

In the early days of digital preservation, communication and advocacy involved blunt statements about the social and economic impact of data loss and obsolescence. As solutions have emerged, so messages have become more subtle.

Advocacy has become increasingly about identifying stakeholders and helping them understand:

- how their choices make digital collections more or less resilient; and
- the benefits they will accrue from the active management of well-formed and accessible digital materials
- the necessity of investment – whether time, money or other resources – and the extent to which it is required to achieve these benefits.

In an institutional setting this means understanding all the agents involved in a digital object lifecycle, helping them to prioritise and support those actions that make and keep collections robust, and discouraging those actions which put collections at risk.

Stakeholder Analysis

Stakeholder analysis starts with gaining a clear understanding of the organisation's digital preservation aims before identifying internal and external stakeholders who can influence those

goals. Having identified them, it is then possible to develop a plan that will convey your aims and engage them in the digital preservation process. Approaching this with a clear methodology in mind will produce the best results and will tie in with a number of other digital preservation activities such as policy and strategy development (see Institutional policies and strategies), creating a business case (see Costs, benefits, impact and business cases) and identifying relevant standards and best practice (see Standards and best practice).

The following steps will help facilitate a thorough analysis of stakeholders:

1. **Identify what you hope to achieve** through your digital preservation activities. This may include lists of the principal collections involved and the main aims and objectives as well as potential benefits (see Business cases, benefits, costs, and impact) that will accrue. This will provide a clear reference and focus for advocacy and can later be tailored to the various audiences that are identified.

2. **Identify the groups and individuals** that can inhibit or enable digital preservation activities. These may be internal or external, and any one stakeholder can have multiple roles. For example, you may identify information technology staff as a key group which may then in turn include an IT Services Manager, Programmers and Support Staff. Some of these may be easily accessible inside your own department, some fall within different line management structures, and some will be entirely external. This means you may need to include other managers or service owners within your stakeholder analysis.

3. **Organise the stakeholder groups and individuals into key audiences** that are in a position to influence your goals and priorities. The audiences chosen will probably reflect the working practices of your organisation and/or your approach to digital preservation, perhaps relating to specific parts of your organisational structure (e.g. Senior Management, IT, Information Managers) or by their role in relation to the digital preservation process (e.g. Funders, Depositors, Users).

4. **Establish solid collaborative relationships with the key stakeholder audiences you have identified** to underpin progress towards the aims established in Step 1. Understanding the needs, priorities and constraints of internal and external stakeholders will yield information that directly informs your planning and improves your understanding of what stakeholders want and need from digital preservation activity. Stakeholders may be constrained by budget and/or legislative boundaries of which it will be valuable to be aware. Conversely, they may also have relevant expertise or resource that can be deployed towards digital preservation activity. In addition, understanding the language and terminology used by stakeholders enhances effective communication strategies and can help avoid difficulties that arise when stakeholders understand a term or concept in divergent ways. The ability to use your stakeholders' language generally helps get colleagues and collaborators to buy into your plans. If key stakeholders have conflicting interests you will need to mediate between them.

5. **Building on this two-way engagement, clearly define the important information to be shared with these audiences** that will help secure their buy-in. This should include:

    a Key messages based on your aims and objectives. These should be simple and direct statements written in plain language so they are easily understood by a wide range of

non-specialist audiences. Ideally they should also be aligned with wider organisational strategies and aims.

b Benefits that stakeholders will accrue from participation in/support of the proposed digital preservation activities. For example an IT manager might want to reduce costs of storage by deleting or de-duplicating redundant storage. A clear digital preservation strategy can help them reduce their storage requirements by distinguishing those collections that must be retained from those that are no longer required

c What will be required of them to ensure success. For example, you may wish to develop clear metadata requirements for depositors; or you may wish to give your IT department estimates for the amounts of storage and bandwidth that will be required and when.

d What barriers/misconceptions about digital preservation you may need to address. For example preservation is often confused with just having back-up copies. You may need to tailor language and terminology to specific audiences. For example certain terms such as "archiving" have different meanings in other sectors such as IT.

6. **Make a plan to engage each of the stakeholder groups** building on your knowledge of their priorities, expertise and limits, and using the various messages previously identified. You may need to use different methodologies for the various groups, tailoring your form of communication to best suit the audience and messages to be conveyed. This may include a range of communications channels including presentations, briefing papers and stakeholder working groups as well as developing a variety of plans and resources such as business cases (see Business cases, benefits, costs, and impact), policies (see Institutional policies and strategies) and risk registers (see Risk and change management).

Digital Preservation in the Media

Digital preservation gets surprisingly little attention in the mainstream media. Reporting of digital preservation tends to fall into two clichés: gloomy stories of data loss and an impending 'digital dark age'; or platitudinous statements about indestructible storage.

The reality is more mundane and more subtle. Practical, detailed and achievable requirements that deliver long term access, such as reported in this Handbook, are less attention-grabbing, but can deliver real benefits to institutions and their user communities.

In some advocacy contexts it may be useful to refer to a common vocabulary to support explanation of key terms and concepts in digital preservation. Some examples are suggested in the resources section below.

The broader digital preservation community has created short animations for advocacy such as those selected in the resources section below. These are short, entertaining, and often helpful in getting key messages about digital preservation across to non-specialist audiences and the general public.

Resources

Team Digital Preservation and Nuclear Disaster: An Animation

https://www.youtube.com/watch?v=pbBa6Oam7-w

Entertaining cartoon on the importance of trusted digital repositories, metadata, and refreshing digital media. (3 mins 18 secs)

Team Digital Preservation and the Aeroplane Disaster

https://www.youtube.com/watch?v=EKnsZZzuUr4

Entertaining cartoon on the effects of obsolescence and importance of migration. (3 mins 37 secs)

Team Digital Preservation and the Arctic Mountain Adventure

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

Entertaining cartoon on the importance of preservation planning. (4 mins 22 secs)

Team Digital Preservation and the Deadly Cryptic Conundrum

https://www.youtube.com/watch?v=Yun9hkPPF9M

Entertaining cartoon on the importance of representation information. (4 mins 9 secs)

Case Studies

Increasing Participation in Internal RDM Training Sessions

http://www.dcc.ac.uk/resources/developing-rdm-services/increasing-participation-training

This case study looks at the approaches taken by two Jisc MRD Projects to ensure good attendance at their internal research data management (RDM) training sessions. 2013 (4 pages).

Defining and Formalizing a Procedure for Archiving the Digital Version of the Schedule of Classes at the University of Michigan

http://files.archivists.org/pubs/CampusCaseStudies/Case2Final.pdf

Nancy Deromedi of the University of Michigan describes forming a partnership with a key administrative unit that had not been to date a receptive partner on campus, and raising the awareness of the archival considerations as the unit transitioned from a hybrid system of digital and paper to a solely digital process. April 2008 (8 pages).

Procurement and Third Party Services

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Introduction

This section provides an overview of key issues and guidance in selecting and using and third-party services for digital preservation. The ways in which a service may be procured often vary according to sector or country. Individual organisations must identify and follow their statutory and regulatory purchasing policies to ensure that services are purchased using the correct procedures. Failure to purchase under the specific guidelines could lead to a serious issue possibly involving compensation to other potential contractors disadvantaged by incorrect purchasing processes.

Three tables are provided as part of the guidance: *Staff resources for procurement tasks*; *Issues and potential advantages and disadvantages of using third party services in digital preservation activities*; and a *Checklist for assessing storage readiness for digital preservation* as procurement is often a major component of implementing archival storage (see Storage and Cloud services). The final Resources section provides additional pointers to and summary description of further guidance and case studies.

Cost will clearly be a key consideration when deciding whether or not to contract out digital preservation but there are also other factors to consider and the advantages and disadvantages of each will need to be balanced against the overall mission of the institution. These include the contract, service level agreement, functionality and quality of the services provided, integration with the institution's processes and environment, disaster recovery and business continuity plans, ability to exit the service if needed, and how the service can be monitored and measured. For example, legal requirements for data privacy or confidentiality may influence whether outsourcing is appropriate or not given the jurisdiction of the service provider and where the service is physically located.

Outsourcing specific tasks or services is by no means a new phenomenon. Repositories have contracted out some of their operations for decades. This is an area in which lessons learned from outsourcing in other services can be of value. A major learning experience which is directly applicable to the digital environment is the critical importance of having sufficient staff resources and knowledge of the technology to be able to prepare effective specifications.

Staff resources for procurement tasks

The extent to which the potential advantages of using third party services can be maximised and the potential disadvantages minimised will be heavily dependent on dedicating staff resources to the following activities:

Staff resources for procurement tasks
Establishing the organizational remit and appropriate governance when selecting third-party services: • Advocate the digital preservation concept. • Involve the appropriate internal stakeholders early in your thinking. • Develop a communications strategy not just for your procurement team but for broader stakeholders. • Maintain an up to date risk register for the procurement. • Use the expertise within your organization: for example do you have a procurement section?
Establishing clear and realistic requirements: • Align business case with identified needs within organisation. This may take some time but is vital to achieving a good outcome. • Create an environment in which your stakeholders can contribute to the discourse and feel they have had input. • Learn from experience of others in digital preservation community as a whole: this could include reference site visits and/or sharing documentation and viewpoints. • Define which preservation functions are to be included, e.g. ingest; storage, preservation planning, curation and management activities. Are all activities to be outsourced to third party provider or just part of the digital preservation framework? • Distinguish between essential functionality and desirable 'added value' functionality: using a particular requirements methodology for example the MoSCoW template provides important discipline not just in the early procurement stages but for any project plan going forwards. • Have unambiguous and measurable requirements that you can use to clearly show the contractor if they are meeting them or underperforming.
Clarifying legal requirements: • Follow institutional and regulatory procurement requirements and processes. • Data Protection, FOI, other sensitive content and copyright will all need to be considered. • It is worth spending time upfront on negotiating your contracts and agreements with third party providers. Misunderstandings in the future are time consuming.

- If you require changes to the contract or agreement offered it is vital to secure them before signing any binding contract in law. Build in review points to the contract and understand what levers you have at your disposal.
- Ensure that you are not obligated to award a contract to provide you with flexibility up to that point.
- Ensure that you can terminate the contract in a minimally disruptive way if the contractor is not meeting the requirements of the contract. Where possible only pay for the goods and services you have received and not those that you may receive in the future.
- Clearly understand what services and products are being offered within the baseline costs of the contract and which will incur additional costs.
- Make all legal requirements, including the legal jurisdiction/governing law, available to potential contractors as early as possible in the procurement as this may greatly affect if they take part in the process.

Maintaining good communication between the contractor and the institution:

- Service Level Agreement to identify roles and responsibilities of each party.
- Access to technological infrastructure only or external staff time / development support also?
- The softer vendor relationship building skills are also important in this context.
- Is there an active user community for your chosen system that provides feedback and good interaction with the contractor?

Undertaking quality assurance checks:

- Establish responsibility for functions such as integrity checking.
- Match quality assurance checks with the measurable requirements you specified in the contract and ensure the supplier is meeting requirements or changing /correcting their practices to meet them.
- Audit / compliance with legal responsibilities.

Developing and monitoring the contract:

- This may seem premature but an exit strategy should be identified upfront. Digital preservation function will outlast commercial service provider and current technological infrastructure.
- Understand your rights regarding your data. Are there costs of retrieving data required for access or transfer to another provider?
- Be mindful of the market and financial models used by vendors. You may need to think outside the box as these models might not match the financial model prevalent in your organisation (capital expenditure vs revenue expenditure is one frequent dilemma).

- Awareness of any changes to technological environment for third party provider.
- Keep up to date with the market after you have concluded your procurement. You need to know how commercially robust your vendor is. Update your due diligence checking periodically.

These costs will need to be added to the overall contract costs when calculating the cost benefit of using third party services for digital preservation, bearing in mind that most of these costs will be or should be incurred even if preservation is not outsourced.

Issues and potential advantages and disadvantages of using third party services in digital preservation activities

Issue	Potential advantage of using 3rd party services	Potential disadvantage of using 3rd party services
Limited staff, skills and experience	• Provides specialist skills and experience which may not be available within the institution.	• Without some practical experience and expertise, it will be difficult to develop and monitor effective contracts. Without practical experience it will also be difficult to understand and communicate effectively the requirements of the organisation (or to assess whether they are technically feasible or not). • Without practical experience it will also be difficult to understand and communicate effectively the requirements of the organisation (or to assess whether they are technically feasible or not)
Costs	• Avoids the need to develop costly infrastructure (particularly important for small institutions). • If there are economies of scale, outsourcing may well be cost effective.	• There is very little established bench marking. It is still too new an area. • Risk of business failure. • Until the market increases there may be an over-dependence on one contractor.
Speed of deployment	• Allows action to be taken in the short to medium term, pending	• Unless there are adequate exit strategies, may be locked into an

	development of infrastructure.	outsourcing contract longer than intended.
Core competencies	• Allows the institution to focus on other aspects of service provision.	• Danger of either not developing or losing specialist skills base. Still need ability to make informed decisions.
Access considerations	• Monitoring usage may be more efficient (assuming the contractor has a demonstrated ability to deliver meaningful usage statistics). • There may be synergies and cost savings in outsourcing access and preservation together.	• • Difficult to control response times which may be unacceptably low and/or more costly, especially for high-use items. • May be difficult to forecast future needs in this area.
Rights Management	• Avoids what is often a resource intensive activity for the institution.	• May significantly increase the cost of the contract and/or complicate negotiations with third party rights holders.
Security	• Contract can guarantee security arrangements required by the institution.	• Lack of control, especially for sensitive material.
Quality control	• A watertight contract will build in stringent quality control requirements.	• Risk of loss or distortion may still be unacceptably high for highly significant and/or sensitive material.
Storage	• Access to professionally managed and experienced storage arrangements with easy replication of content and integrity checking.	• Issues of trust and legal considerations when storing sensitive data. • Difficult to anticipate the actual costs of some services e.g. cloud storage and computing because the organisation often does not know exactly how much service it will need. This is uncertainty can be reduced with experience.

Checklists for selecting and comparing service providers

Checklists and standards can be valuable starting points when considering or evaluating the use of third-party services as they are ready made lists that you can easily adopt or adapt to fit your needs. In particular, checklists help you identify things that you might otherwise forget to consider as well as helping you to express issues and requirements clearly.

Checklists work well when coupled to a maturity model. For example, the NDSA preservation levels allow a checklist to be constructed to see how well a service provider delivers to each level. An organisation identifies what level of maturity they need both now and in the future and then looks for service providers with matching levels.

Checklists and standards for repository services are valuable starting points because you can pick and choose the parts of the checklist that would apply to the specific services you seek. Examples of relevant checklists and standards are available in Resources and are also discussed in more detail in the Audit and certification section of the Handbook.

A Handbook checklist for assessing storage readiness for digital preservation is provided below:

Checklist: questions for your preservation storage service provider
☐ What level of redundancy does the storage system provide? How many physical locations is digital material held in? What is the geographical distance between them?
☐ Are different types of storage technology employed to mitigate/spread risk? For example online and off-line storage.
☐ If a file has become corrupted or unintentionally altered, how does this get detected and when does detection happen? Are audit trails or other forms of logging available to show that data integrity checks have been done and to show the result?
☐ What is the disaster recovery strategy, for example if a storage system fails or there is a natural disaster at a storage site then how are digital materials recovered? When was the last time this DR strategy was tested?
☐ What is the storage migration strategy to address technical obsolescence? What happens when the system is at the end of its life and content needs to be migrated to a new system? Is the content still accessible during this process?
☐ What is the exit strategy when using a given type of storage (e.g. onsite, cloud) for example what happens if the vendor of the storage system goes out of business?

- [ ] What measures are in place to contain corrupted or altered files, for example quarantining files to prevent them from being replicated?

- [ ] What security and auditing measures are in place to prevent unwanted access and/or modification of the digital materials?

- [ ] Who is responsible for monitoring and managing the storage system to ensure it is functioning correctly? Is there continuity of staff in cases of holiday, sickness or departures?

- [ ] What contracts, warranties or guarantees come with the storage solution or service that commit the vendor or supplier to support, recovery or replacement if there are any problems?

- [ ] What approach or support is in place for storage technology watch and risk assessment so that migrations, refreshes, upgrades or maintenance can be planned and executed in a timely way?

- [ ] Are the costs and risks clear so that a trade-off can be assessed and made between number of copies, type of storage, ease of access, and safety of the digital materials?

- [ ] What standards does the provider aim to comply with? (e.g. OAIS, Information Security Standards) Does it aim to achieve recognition as a trusted digital repository?

- [ ] How can the provider demonstrate they are doing what you have agreed?

Resources

OAIS: Open Archival Information Systems: Reference Model for an Open Archival Information System. Recommended practice

http://public.ccsds.org/publications/archive/650x0m2.pdf

Provides a useful shared terminology and functional model when identifying requirements for procuring third party digital preservation services. (135 pages).

Data Seal of Approval (DSA)

http://datasealofapproval.org/en/information/guidelines/

The Data Seal of Approval is a self-assessment process for digital archives, aimed specifically at those archives that hold data. This repository assessment includes a 16 point checklist.

ISO16363: 2012 Audit and certification of trustworthy digital repositories

http://www.iso16363.org/

ISO 16363 is an evidence-based audit framework for digital preservation consisting of more than 80 criteria that can be used for self-audit or external audit. The criteria used in the standard look across the entire organisation and not just the technical system in which collection content is stored. The CCSDS Magenta Book pre-print version of the standard is freely available at http://public.ccsds.org/publications/archive/652x0m1.pdf.

DIN 31644 Information and documentation - Criteria for trustworthy digital archives

http://files.dnb.de/nestor/materialien/nestor_mat_17_eng.pdf

The extended self-assessment process for digital archives is a helpful checklist developed by nestor on the basis of the DIN 31644 Information and documentation - Criteria for trustworthy digital archives standard.(44 pages).

The NDSA Levels of Digital Preservation: An Explanation and Uses

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

The US National Digital Stewardship Alliance (NDSA) Preservation Levels are used widely throughout the Handbook and are helpful in thinking about many areas of digital preservation. There are also Mappings of NDSA preservation levels to cloud storage vendor profiles by AVPreserve.(7 pages).

Where to keep research data DCC Checklist for Evaluating Data Repositories

http://www.dcc.ac.uk/sites/default/files/documents/publications/Where%20to%20keep%20research%20data.pdf

A useful Digital Curation Centre checklist on where to keep research data safe that includes Service Level Agreement maturity levels. It is mainly concerned with external third-party repositories that offer a managed service to the UK research community.(20 pages).

The National Archives Cloud Storage Guidance

http://www.nationalarchives.gov.uk/archives-sector/digital-collections.htm

Provides information about procurement in the context of cloud computing services for preservation purposes, including case studies from several institutions (see below). It is particularly notable for its consideration of the legal issues.

DPC procuring preservation event

http://www.dpconline.org/events/previous-events/1150-procuring-preservation-writing-and-understanding-requirements-in-digital-preservation

For an overview of some of the elements of scoping requirements see the individual presentations listed. Presentations on Requirements analysis, and Procuring Preservation: hoops, hurdles and processes are particularly relevant.

Case studies

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).

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).

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).

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).

Partnering with IT to Identify a Commercial Tool for Capturing Archival E-mail of University Executives at the University of Michigan

http://files.archivists.org/pubs/CampusCaseStudies/CASE-14-FINAL.pdf

Aprille Cooke McKay, Bentley Historical Library, University of Michigan, examines the challenges and opportunities of partnering with IT to issue a Request for Proposal (RFP) for commercial e-mail archiving software. 2013 (53 pages).

University of Sheffield Procurement Case Study

https://www.sheffield.ac.uk/library/special/speccoll

A summary of the process of procuring a digital preservation system at the University of Sheffield. (2 pages).

Audit and certification

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Introduction

Organizations are increasingly interested in evaluating their digital preservation infrastructures against an assessment framework, and audit, certification, and self-assessment are hot topics in digital preservation. It is worth taking a moment to consider the difference between a self-assessment exercise and an audit.

Audit and certification is a formal process commonly carried out and delivered by external service providers. It is often a time consuming experience with exactingly high requirements that demonstrate to an external audience that a particular standard is being complied with.

Self-assessment is a precursor, or alternative, to a full audit and is typically delivered by staff inside of the organization, and the results are usually of highest value to the organization being assessed (rather than an external audience). Self-assessments can be useful in identifying practices which are underdeveloped and require improvement, particularly if an organization is interested in pursuing full audit and certification at a later date.

Many of the benefits can be summarised as ensuring that a repository can be trusted. The concept of a trusted or trustworthy digital repository is now broadly recognised in the digital preservation community. The following section summarises the work that has taken place over the past 10 - 15 years to get us to this point.

Background to development of audit and certification frameworks

Audit and certification methods for digital preservation implementations have been in development for well over a decade with different organizations developing different methodologies in parallel. In Europe these are now coalescing under the European Framework for Audit and Certification of Digital Repositories.

The OAIS Reference Model (ISO, 2012a) (see Standards and best practice) influenced the development of the different methodologies, which began with the publication of Trusted digital repositories: Attributes and responsibilities (RLG/OCLC, 2002). This was refined as the draft publication An audit checklist for the certification of trusted digital repositories (RLG-NARA, 2005) before being finalised as TRAC (Trustworthy Repositories Audit & Certification: Criteria and Checklist) (CRL, 2007).

Equivalent activity was also taking place in both the Netherlands and Germany. The self-assessment process, Data Seal of Approval developed by DANS (Data Archiving and Networked Services), was released in 2008. Meanwhile, based on recommendations from a working group of nestor, the German Standards Committee (DIN) adopted *DIN 31644 Information and documentation - Criteria for trustworthy digital archives*.

Following their publication of the OAIS standard, and the later adoption of OAIS as an ISO Standard, in September 2011 the Consultative Committee for Space Data Systems released recommended practice on "Audit and certification of trustworthy digital repositories", This was subsequently adopted and published as *ISO 16363 2012 Audit and certification of trustworthy digital repositories* (ISO,2012b).

Current assessment options and the European Framework for Audit and Certification

The apparent proliferation of repository audit standards has been frequently cited as a barrier to participation. Consequently the European Commission has hosted a series of meetings to discuss a European-wide approach, and there is now a Memorandum of Understanding to define a European Framework for Audit and Certification of Digital Repositories. This memorandum effectively creates a tiered approach to certification, allowing an entry-level self-assessment and peer review based on the Data Seal of Approval, a more extensive self- assessment (based on DIN 31644 or ISO 16363), and a full scale external audit based on ISO 16363.

1. Data Seal of Approval

The Data Seal of Approval (DSA, 2008) is a self-assessment process for digital archives, aimed specifically at those archives that hold data. Though an outlay of time is needed to apply for the DSA,

it is far less onerous than ISO 16363, having only sixteen guidelines on which the organisation is assessed. The guidelines are based on the following five criteria:

- The data can be found on the Internet;
- The data are accessible (clear rights and licences);
- The data are in a usable format;
- The data are reliable;
- The data are identified in a unique and persistent way so that they can be referred to.

Though the DSA is on the surface a self-audit, this self-audit is then peer reviewed before a seal is awarded, thus adding a level of authority to the process. Openness and transparency are encouraged and institutions are asked to make their evidence (essentially documentation, policies and procedures) freely available online. Unlike an audit under ISO 16363, the peer reviewer is not required to visit the institution to see that the policies and procedures are working in practice, so this process is very much based on trust.

DSA are in the final stages of reviewing proposed amendments to the DSA Guidelines as a result of work with the World Data System through the Research Data Alliance. Details of when and how the transition to new guidelines will be managed will be released in due course, but in the meantime the current seal will be extended through 2017.

2. DIN 31644 Information and documentation - Criteria for trustworthy digital archives

The DIN Standards Committee in Germany adopted *DIN 31644 Information and documentation - Criteria for trustworthy digital archives* based on recommendations from a working group of the German competence network for digital preservation (nestor). The standard consists of requirements for a trustworthy digital repository structured in three sections:

The **organisational framework** requires that:

- The repository has defined goals for the selection of digital material and accepts the responsibility to preserve them over the long- term;
- The repository has a defined community for whom access and the ability to interpret digital materials will be provided;
- There is observation of legal and contractual rules between data creators and the digital repository;
- Sufficient organizational structures are provided in terms of personnel, finance, long-term planning and continuity of service;
- Processes and responsibilities are defined and documented.

Object management requires that:

- The integrity and authenticity of digital material are maintained;
- A strategic plan for digital preservation activities is in place;
- Information packages for ingest, storage and dissemination are defined;

- Adequate documentation is provided including permanent identifiers and sufficient structural, technical, rights and change metadata;
- The digital material and related metadata are packaged together for permanent preservation.

Infrastructure and security requires that:

- The IT infrastructure can deal with the digital material adequately and is secure.

DIN 31644 is in German but an English translation is provided by nestor on their website.

The extended certification process undertaken by nestor takes about three months. Guidance on this process, (nestor Certification Working Group, 2013) is available on their website. This certification process should not be confused with full external audit- this requires formal accreditation under ISO 16363.

3. ISO 16363 Audit and certification of trustworthy digital repositories

ISO 16363 is an evidence-based audit framework that uses the term 'repository' to mean the organisation responsible for digital preservation rather than just the technical infrastructure being used for storage. The criteria used in the standard look across the entire organisation and not just the technical system in which collection content is stored. Metrics are grouped into three areas:

- **Organizational Infrastructure**: including governance, organizational structure, staffing, procedural accountability, policy framework, financial sustainability and contracts, licensing and liabilities;
- **Digital Object Management**: including acquisition and ingest, preservation planning, creation and preservation of Archival Information Packages (AIPs), and information and access management;
- **Infrastructure and Security Risk Management**: including technical infrastructure, risk management and security risk management.

Terminology used in ISO 16363 is directly aligned with that of OAIS and the standard asks directly about both OAIS information packages and functional areas. A basic understanding of OAIS is therefore useful for those seeking to understand ISO 16363 and deliver an assessment against it.

With over 100 metrics spread across the three areas, undertaking an ISO 16363 audit or assessment is a significant commitment similar to many other ISO standards applied across organisations. A relatively small number of organisations have utilised the ISO 16363 standard since it was published. Some have sought certification by external auditors whilst others have undertaken self-assessments. Houghton (2015) acknowledges that even though a self-assessment is not an audit it is nonetheless a significant undertaking that should be tailored to organisational circumstances.

ISO 16363 follows ISO practice for certification which assumes that those carrying out the audit are themselves certified. Two other ISO standards support this:

- *ISO 16919 Requirements for bodies providing audit and certification of candidate trustworthy digital repositories* (ISO, 2011) that sets out the requirements for any organisation that certifies the auditors for ISO 16363; and
- *ISO 17021 Requirements for bodies providing audit and certification of management systems* (ISO, 2012a) provides a mechanism to audit accreditation bodies.

An agency called PTAB (Primary Trustworthy Digital Repository Authorisation Body) offers training for auditors and those preparing for audit. Other agencies including the Center for Research Libraries are also providing audits against these standards.

4. Other frameworks and tools for self-assessment

A useful entry level resource is the Levels of Digital Preservation from NDSA (NDSA, 2013). This is particularly useful for those institutions that are just starting on starting out and can be used to benchmark initial steps. The NDSA levels are used extensively in the Handbook (see Getting started, Fixity and checksums, Information security, and Storage). Risk assessment frameworks and tools can also contribute to audit assessments (see Risk and change management).

Which audit or assessment option should I choose?

The 2010 Memorandum of Understanding described above, effectively identifies a tiered approach to certification. The amount of effort required for each level increases, though so does the formality of the output. The choice of assessment framework for any given organisation should therefore take at least the following into consideration:

Selecting an assessment framework	
What do you want to achieve from your audit?	What level of trust are you trying to engender? Do you seek certification from an external authority, or is self-assessment sufficient?
How much effort or funding is available to deliver the assessment?	ISO 16363 is a large undertaking that requires a significant amount of effort to gather the available evidence and run the audit; DSA has far fewer metrics and can be completed in a much shorter period. DIN 31644 has two assessment options with varying effort needed.
What type of content does your institution hold?	To date, DSA has been specifically developed for data-holding repositories, while DIN and ISO 16363 are both content-type neutral.
What framework, if any, will carry most weight in your organization or with your external stakeholders?	Is there any national preference for a framework or a framework commonly used by similar organisations that you should use?

The choice of assessment framework should not be made independently and can often be directly influenced by the value that an assessment may have for other parts of the organization. Discussing the options with organizational peers and managers can be a useful first step in ensuring the right option is selected and support is secured from other areas of the organization from the outset.

How to run an audit or self-assessment

Once an appropriate methodology has been selected, a straightforward way to proceed is to develop the initiative as a project and produce a project plan. Advice on project planning is prolific online and you should consult this if your organization does not have an agreed process for project management. If it does have a process, then you should become familiar with it and plan your project using this methodology (or secure the assistance of a local project manager). Your plan should include at least the following sections:

- Scope: What content is in scope of the assessment?
- Timeframe: When will the assessment take place and when will it deliver results?

- Stakeholders: Who will deliver the assessment? Who else needs to be interviewed or consulted?
- Governance: Which group will have governance of the assessment and results?
- Communications: How will the process and outcome be communicated to stakeholders?
- Next steps: How will the results be implemented?

If you are running an ISO 16363 assessment you should consult the advice on the ISO 16363 Primary Trustworthy Digital Repository Authority Body website. The audit preparation page outlines the steps that should be taken when running a full audit and these can be adapted for a self-assessment. Similarly, the Data Seal of Approval website has an online self-assessment tool that will guide you through an assessment. PDF or HTML versions of the assessment manual guidelines are also available.

Resources

APARSEN Report on Peer Review of Digital Repositories

http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2012/04/APARSEN-REP-D33_1B-01-1_0.pdf

Lessons learnt to date from the process of repository certification have been usefully summarized by the APARSEN project in this report. It suggests although there has been considerable progress, arguably audit procedures are not yet fully bedded down and some issues remain for both auditors and repositories. (2012, 50 pages).

Digital Preservation Management tools: Principles

http://dpworkshop.org/workshops/management-tools/principles

For organizations that are committed to becoming a Trusted Digital Repositories (TDR), a formative step for developing a sustainable digital preservation and curation program is to adapt and adopt a set of standards-based principles as a foundation. The principles provide a frame for your program and adopting them is a positive (and hopefully easy) place to start.

Digital Preservation Management tools: Model document

http://dpworkshop.org/workshops/management-tools/policy-framework

Every Trusted Digital Repository needs to have a high-level policy document that explicitly states the scope, purpose, objectives, operating principles, and context of the organization's digital curation and preservation program. The DPM workshop team developed this model document to help

organizations meet this objective. A model document identifies the recommended sections of a digital preservation policy framework with descriptions and examples for each section.

Digital Preservation Management tools: Self-assessment and peer review audit

http://dpworkshop.org/workshops/management-tools/self-assessment

TRAC (Trustworthy Repository Audit and Certification) Review tool developed for the DPM workshop.

The Open Archival Information System (OAIS) Reference Model: Introductory Guide (2nd Edition)

http://dx.doi.org/10.7207/twr14-02

This DPC Technology Watch Report from 2014 provides an accessible short guide to the OAIS standard. Terminology used in ISO 16363 is directly aligned with that of OAIS. The report will help provide a basic understanding of OAIS useful for understanding ISO 16363 and deliver an assessment against it.

Digital Preservation Capability Maturity Model (DPCMM)

https://lib.stanford.edu/files/pasig-jan2012/12F2%20Digital%20Preservation%20Capability%20Maturity%20Model%20in%20Action.pdf

This presentation describes a Digital Preservation Capability Maturity Model (DPCMM) that employs performance metrics based on specifications of ISO 14721 (ISO, 2012a),TRAC, and other good practices. (25 pages).

Data Seal of Approval

http://www.datasealofapproval.org/

In addition to the ISO standards developed by CCSDS, other formal initiatives in this area of archive certification have been the Data Seal of Approval (DSA), and the German Standard on Trustworthy Archive Certification DIN 31644.

European Framework for Audit and Certification of Digital Repositories

http://www.trusteddigitalrepository.eu/Site/Welcome.html

In 2010, the European Framework for Audit and Certification of Digital Repositories was established as a collaboration between the Data Seal of Approval (DSA) certification, the Repository Audit and Certification Working Group of the CCSDS, and the German Standards (DIN 31644) Working Group on Trustworthy Archives Certification. It aims to support an integrated framework for auditing and certifying digital repositories consisting of a sequence of three levels, in increasing trustworthiness.

Case Studies

Preserving the H-Net Academic Electronic Mail Lists

http://files.archivists.org/pubs/CampusCaseStudies/Case11Final.pdf

Lisa M. Schmidt, Michigan State University, describes assessing the existing state of preservation for the H-Net e-mail lists using digital preservation theory and the Trusted Repositories Audit & Certification: Criteria and Checklist (TRAC) evaluation tool. Making recommendations and overseeing the implementation of improvements to make H-Net a trusted digital repository. Ensuring authenticity is the primary preservation issue. 2009 (15 pages).

ADS and the Data Seal of Approval – case study for the DCC

http://www.dcc.ac.uk/resources/case-studies/ads-dsa

Archaeology Data Service colleagues Jenny Mitcham and Catherine Hardman describe the ADS experience in applying for the Data Seal of Approval (DSA). They identify practical information about the DSA application process. They also outline issues ADS faced in undertaking the process and the potential benefits they envisage from DSA self-certification. 2011.

Self-assessment of the Digital Repository at the State and University Library, Denmark – a Case Study

https://ipres-conference.org/ipres14/sites/default/files/upload/iPres-Proceedings-final.pdf

In this iPres 2014 paper, the authors describe the process and the benefits of performing an audit based on self-assessment and ISO 16363 for the digital repository of the State and University Library in Denmark. (p.272-279 of 385).

TRAC Audit: Lessons

http://blog.dshr.org/2014/08/trac-audit-lessons.html

This is the third in a series of blog posts by David Rosenthal about CRL's TRAC audit of the CLOCKSS Archive. Previous posts announced the release of the certification report, and recounted the audit process. This post look at the lessons CLOCKSS and others can learn from their experiences during the audit.

Trustworthiness: Self-assessment of an Institutional Repository against ISO 16363-2012

http://www.dlib.org/dlib/march15/houghton/03houghton.print.html

In 2013, Deakin University Library undertook a self-assessment against the ISO 16363 criteria. This experience culminated in the current report, which provides an appraisal of ISO 16363, the assessment process, and advice for others considering embarking on a similar venture.

Managing an ISO 16363 Self-Assessment: A How-To Guide

http://www.dcc.ac.uk/sites/default/files/documents/IDCC16/18_Managing_ISO16363.pdf

A short poster presented at the International Digital Curation Conference (IDCC) in 2016 by Maureen Pennock and Caylin Smith of the British Library.

References

CRL, 2007. *Trustworthy Repositories Audit & Certification: Criteria and Checklist*. Available: http://www.crl.edu/sites/default/files/d6/attachments/pages/trac_0.pdf

DIN, 2012, *DIN 31644 Information and documentation – Criteria for Trusted Digital Repositories*. Available: http://www.nabd.din.de/cmd?level=tpl-art-detailansicht&committeeid=54738855&artid=147058907&languageid=de&bcrumblevel=3&subcommitteeid=112656173

Houghton, B., 2015. *Trustworthiness: Self-assessment of an institutional repository against ISO 16363-2012*. DLib Magazine, 21(3/4). Available: http://www.dlib.org/dlib/march15/houghton/03houghton.html

ISO, 2011. *ISO 16919:2011 - Space data and information transfer systems - Requirements for bodies providing audit and certification of candidate trustworthy digital repositories*. Available: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=57950

ISO, 2012a. *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

ISO, 2012b. *ISO 16363:2012 - Space data and information transfer systems – Audit and certification of trustworthy digital repositories*. Geneva: International Organization for Standardization. Available: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=56510

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

nestor Certification Working Group, 2013. *Explanatory notes on the nestor Seal for Trustworthy Digital Archives*, nestor Materials 17, July 2013. Available: http://files.dnb.de/nestor/materialien/nestor_mat_17_eng.pdf

RLG/OCLC Working Group on Digital Archive Attributes, 2002. *Trusted digital repositories: Attributes and responsibilities*, Mountain View, California. Available: http://www.oclc.org/content/dam/research/activities/trustedrep/repositories.pdf

RLG-NARA Task Force on Digital Repository Certification, 2005. *An audit checklist for the certification of trusted digital repositories*, Mountain View. Available: https://web.archive.org/web/20051126181100/http://www.rlg.org/en/pdfs/rlgnara-repositorieschecklist.pdf

Legal compliance

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Introduction

The information provided in this section is intended solely as general guidance on the legal issues arising from various aspects of digital archiving and preservation and is not legal advice. It does not attempt to provide guidance on general legal issues which impact on the operations of libraries, archives and other repositories, as these are covered in a number of other reference works. It is written from a UK perspective and legislation in this area will vary from country to country. Although it principally covers UK and European legal issues, many of the topics will also apply in general terms to other jurisdictions.

An adviser–client relationship is not created by the information provided. If you need more details pertaining to your rights and obligations, or legal advice about what action to take, please contact a legal adviser or solicitor.

Legal issues

'those engaged in digital preservation must work within the law as it stands. This requires both a good general knowledge of what the law is, and a degree of pragmatism in its application to preservation work. Such knowledge enables the archivist to avoid the pitfalls of over-cautiousness and undue risk aversion, and to more accurately assess the risks and benefits of taking on the preservation of new iterations of digital work.' (Charlesworth, 2012, p.3)

Intellectual property rights (IPR) and preservation

Intellectual Property Rights (IPR) are a section of UK law that include patents, trademarks, copyright and associated rights - such as the moral rights of the author (see Stakeholders, contract and grant conditions) and performance rights. The preservation of digital materials often requires the use of on a range of strategies, and this creates IPR issues that are arguably more complex and significant for digital materials than for analogue media. If not addressed these can impede or even prevent preservation activities.

What is different about copyright and digital materials?

Among the range of IPRs, copyright has a specific importance when considering digital preservation actions. UK copyright law was developed with analogue material in mind. Traditional analogue materials are relatively stable, and well established legal and organisational frameworks for preservation are in place. The legal framework for undertaking preservation work on digital material

is not as well developed and good preservation practices are not always recognised, or allowed for, by existing provisions in current legislation.

Copyright makes a distinction between ownership of the physical manifestation of a work, such as a book or work of art, and the separate right to reproduce it (the right to copy). Digital material by nature does not align to this distinction and can cause confusion when applied in the field. In the case of digital material, core repository practices such as providing access to users and routine preservation activities, often involve the deliberate or inadvertent creation of copies. Without appropriate rights clearance, licences or statutory exceptions these copies may constitute copyright infringements. Digital material therefore poses a different set of considerations for repositories holding this category of content. In addition, unlike physical material, digital material requires consideration of dependencies such as hardware and software which all have their own separate intellectual property considerations.

A second significant difference is the relatively short commercial and technological lifespan of digital material. The duration of IPR in digital materials extends beyond both commercial 'shelf life' and in almost every case the technology on which they depend. This forms a three-fold issue, in terms of procuring licenses to replicate content, licences for software to access content, and rights clearance of "abandoned" digital material, in addition to the added urgency of undertaking these actions.

Copyright exceptions

Although the Copyrights, Design and Patent Act (1988) limits possible preservation actions for digital material, exceptions for archives, libraries and museums have been introduced to address the unique requirements for preserving it. From a preservation perspective the most important provision (Intellectual Property Office, 2014) is the right to produce any number of copies required for the purpose of preserving digital material. Another important exception is the dedicated terminal exception, which enables a digital copy (i.e. one of the copies created under the preservation exception), to be made available on a dedicated terminal accessible to walk in users. These provisions only extend to items held permanently in the collection. The exceptions enables those institutions covered by the exemptions to hold copies of material in various file formats and thereby adhere to what is considered good preservation practice while staying within the law. Note that the copyright exception provisions do not overrule the moral rights of the author which must still be considered when undertaking preservation work.

The exemption to copy for preservation does not however take into account the dependent nature of digital material, and third party software dependencies can still form a barrier to preservation actions. This is particularly an issue observed in preservation strategies which rely heavily on retaining the wider technical environment of the digital objects in question. For example, emulation as a preservation strategy requires use of original operating systems and software external to the repository's permanent collection (see Preservation action). It is important to consider the additional costs and time of maintaining relationships with third party rights holders that follow from dependencies not covered in the exceptions.

Orphan works

For institutions looking to publish digital surrogates of analogue material, the Orphan Works Licensing Scheme run by the UK's Intellectual Property Office, as well as the EU Orphan Works Exception are likely to impact digitisation work and planning. The Licensing Scheme allows for both commercial and (in the case of heritage institutions) non-commercial digitisation of any type of material in which it has not been possible to trace the rights holders of the material following a 'diligent search'. The licence is a pay scheme limited to a seven year period and for use exclusively in

the UK. Repositories need to plan and budget for renewal of such licenses. The EU Orphan Works Exception, on the other hand, is restricted to text based and audio visual works only (and artistic works as long as they are embedded in the former), and museums, libraries, archives, educational establishments and public broadcasters. Here, the benefit is that the diligent searches are self certified and the preservation copy of the work, created under the preservation exception, can be placed on line for non commercial uses, for example, thus assisting greatly with digitisation activities.

Access and security

Some of the additional complexity in copyright issues relates to the fact that digital materials are also easily copied and re-distributed. Rights holders are therefore particularly concerned with controlling access and potential infringements of copyright. Digital Rights Management technologies (DRM) developed to address these concerns and provide copyright measures, such as copy protection software for files and intentional physical errors to CD/DVDs, can inhibit or prevent actions needed for preservation. DRM technologies are also in themselves subject to obsolescence. These concerns over access and infringement need to be understood by organisations preserving digital materials when negotiating deposit agreements with rights holders, and addressed by both parties in negotiating rights and procedures for preservation. Having clear deposit procedures in place can mitigate future access issues (See Negotiating rights).

Web archives and legal deposit

The legal status of web archives and processes of electronic legal deposit vary from country to country: some governments have passed legal deposit legislation but restrict access solely to library reading rooms. In others there is no legal deposit legislation and collections are either built solely on a selective and permissions basis or are held in 'dark archives' that are inaccessible to the public. In the UK, legal deposit libraries have the right to gather and provide access to copies of all websites published in the UK domain. However, access to the collection is restricted to library reading rooms (See Milligan, 2015). Parallel to this, Web Archives maintained by The National Archives (UK) operate with a smaller scope relating to government publications and clearer statutory powers derived from public records legislation (see Other statutory requirements).

The US-based Internet Archive, probably the largest and most used web archive, has no explicit legislative permission to harvest websites or to publish them. It operates on a 'silence is consent' approach, deleting from their collections any websites that an owner requests to be removed. In contrast, the Library of Congress operates on a permission basis meaning that they have to seek explicit approval from copyright holders before harvesting or publishing their content.

Other statutory requirements

Other statutory requirements may also apply and influence preservation of digital resources.

The requirements of public records legislation and the related expectations of the Freedom of Information Act apply to government records including those in digital form. Statutory and regulatory retention periods apply to many digital records (e.g. for accounting and tax purposes). Although these are often of limited duration, it is notable that requirements for retention of digital records in some sectors (e.g. the pharmaceutical industry, social care and health records), are of increasingly long duration. In such cases long-term preservation strategies will apply as technological change will almost certainly affect access to such records.

Information may be subject to data protection laws and relevant privacy legislation protecting information held on individuals. In the UK, the Information Commissioner's Office oversees adherence to data protection and privacy issues.

Information can also be subject to confidentiality agreements. Privacy and confidentiality concerns may impact on how digital materials can be managed within the repository or by third parties, and made accessible for use. Data protection law also impacts on data movement outside of Europe - an important consideration for organisations investing in server space abroad.

EU rulings on an individual's right to have their personal information removed from Internet search engines in certain circumstances has a significant impact on the practices of organizations working with digital content sourced from the web (Koops, 2011). The obligation to avoid doing harm to individuals when saving their data over long periods of time is reflected in the principle of the right to be forgotten, through the implementation of Article 12 of Direction 95/46/EC in the case law of multiple European nations.

Stakeholders, contract and grant conditions

Some digital materials are the result of substantial financial investment by public funds (e.g. research councils) and/or publishers, and intellectual investment by individual scholars and authors. Each of these stakeholders may have an interest in preservation; the organisation preserving these will need to acquire permissions from them to safeguard and maximise the financial investment or the intellectual and cultural value of the work for future generations. Such interests could be manifested through contract, licence, and grant conditions or through statutory provision such as "moral rights" for the authors.

Investment in deposited materials by the repository

Holders of the material over many decades will almost certainly need to invest resources to generate revised documentation and metadata and generate new forms of the material if access is to be maintained. Additional IPR issues in this new investment need to be anticipated and future re-use of such materials considered. Where a depositor or licensor retains the right to withdraw materials from the archive and significant investment could be anticipated in these materials over time by the holding institution, withdrawal fees to compensate for any investment may be built into deposit agreements (See Negotiating rights).

Rights management

As outlined in Legal issues, it is important that licensing issues, copyright and any other intellectual property rights in digital resources to be preserved, are clearly identified and access conditions agreed with the depositor and/or rights holders. If the legal ownership of these rights is unclear or excessively fragmented it may be impractical to preserve the materials and for users to access them. Rights management should therefore be addressed as part of collection development and accession procedures and be built in to institutional strategies for preservation. The degree of control or scope for negotiation that institutions will have over rights will vary but in most cases institutional strategies in this area will help guide operational procedures. It will also be a crucial component of any preservation metadata (see Metadata and documentation) and access arrangements (see Access).

Negotiating rights

As the volume of digital materials grows and the complexity of rights and number of rights holders in digital media continues to expand, ad hoc negotiation between preservation agencies and depositors and between rights holders themselves becomes more onerous and less efficient. This is

particularly problematic for any UK organisations or activities not covered by the new copyright exceptions.

Development of model letters for staff clearing rights, model deposit agreements, and model licences and clauses covering preservation related activities helps to streamline and simplify such negotiations. Institutions should seek assistance from a legal advisor in drafting such models and providing guidance for staff on implementation or permissible variations in negotiations with rights holders.

A number of institutions have developed models which can be adopted or adapted for specific institutions and requirements. The procedures outlined below are a synthesis of current good practice.

Recommended procedures

- Use legal guidance to frame your rights management policy and to develop documents.
- Develop model letters for rights clearance, model deposit agreements, model licences and clauses for preservation activities.
- If you are licensing material from third parties ensure they have addressed future access to subscribed material in the licence and have robust procedures to support this.
- Prepare reasoned arguments and explanations for your preservation activities suitable for external stakeholders such as rights holders who will need to be convinced of the need, and persuaded that their interests will be safeguarded. Remember their awareness of preservation issues may be low.
- Keep detailed records of rights negotiations. Make a schedule clearly identifying a list of materials deposited and covered by the licence. This will ensure that all that is believed to have been sent by the depositor has been received and may form the basis of an acknowledgement of receipt.
- Preserve information about rights and permissions for all your digital materials. Treat licences, schedules, and rights correspondence as key institutional records to be retained in fireproof and secure environments.

Summary of issues for licences and deposit agreements

The following provides a brief checklist and summary of legal issues which may need to be considered in relation to licences for preservation or deposit agreements for digital materials. Requirements will differ between institutions, sectors and countries and the list should be adapted to individual requirements. This list does not constitute legal advice and you must seek legal counsel for your specific circumstances.

IPR and digital preservation

A clause should be drafted to cover the following:

- Permissions needed for content.
- Permissions needed for associated software.
- Permissions needed for copying for the purposes of preservation. (A section which is applicable for material or organisations not covered by the copyright exceptions)

- Permissions needed for future migration of content to new formats for the purposes of preservation. (A section which is applicable for material or organisations not covered by the copyright exceptions)
- Permissions needed for emulation for the purposes of preservation.
- Permissions in respect of copyright protection mechanisms.

Access

- Permissions and conditions in respect of access to the material.

Statutory and contractual issues

- Statutory permissions and legal deposit obligations in respect of digital materials.
- Grant and contractual obligations in respect of digital materials.
- Conditions, rights and appropriate interests of authors, publishers and other funders.
- Confidential information and protection of the confidentiality of individuals and institutions.
- Protecting the integrity and reputation of data creators or other stakeholders.

Investment by the preservation agency

- IPR in any value added by the preservation agency.
- Withdrawal clauses (and associated fees).

Resources

UK legislation is regularly amended. To ensure that you are accessing the latest updates please refer directly to: http://www.legislation.gov.uk/

Exceptions to Copyright: Libraries, Archives and Museums

https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/375956/Libraries_Archives_and_Museums.pdf

This guidance leaflet published by the Intellectual Property Office in 2014 sets out the exceptions applicable to libraries, archives and museums. It is relevant to anyone who works in or with libraries,

archives or museums in the UK, or copyright owners whose content is held by such institutions. In covers two significant changes in UK law which affect libraries, archives and museums. The first relates to making copies of works to preserve them for future generations. The second allows greater freedom to copy works for those carrying out non-commercial research and private study.

Intellectual Property Rights for Digital Preservation

http://dx.doi.org/10.7207/twr12-02

This DPC technology watch report was published by Andrew Charlesworth in 2012. The document does not cover recent legislation (such as The Legal Deposit Libraries (Non-Print Works) Regulations 2013 and the 2014 Copyright Exceptions for Libraries, Archives and Museums), but is otherwise a relevant introductory work. The report is aimed primarily at depositors, archivists and researchers/re-users of digital works. Intellectual property law, represented principally by copyright and its related rights, has been by far the most dominant, and often intractable, legal influence on digital preservation. It is essential for those engaging in digital preservation to be able to identify and implement practical and pragmatic strategies for handling legal risks relating to intellectual property rights in the pursuit of preservation and access objectives. (54 pages).

Aligning National Approaches to Digital Preservation

https://educopia.org/sites/educopia.org/files/publications/Aligning_National_Approaches_to_Digital_Preservation.pdf

These 2012 Proceedings include two papers on legal issues: Legal Alignment by Adrienne Muir, Dwayne Buttler, and Wilma Mossink (pgs 43-74); and Legal Deposit and Web Archiving by Adrienne Muir (pgs 75-88). The focus of the first paper is on the key issues of legal deposit, copyright exceptions for preservation and access, and multi-partner and cross-border working and rights management; the second paper discusses the challenges of adapting legal deposit a mechanism designed for print publishing to the digital environment. National approaches to key elements of legal deposit framework and the legal issues arising from non-statutory approaches to collecting digital publications for long-term preservation are identified.(342 pages).

Cloud Storage Guidance Appendix Table 3 - Legal Issues

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

Table 3 provided in section 7 as an appendix to the TNA Cloud Storage Guidance published in 2015, lists legal points in greater detail for each of the three key categories:

- Any legal requirements in terms of management, preservation, and access placed upon archives and their parent organisations, by their donors and funders via contracts and agreements or via legislation by Government (e.g. accessibility, availability, information security, retention, audit and compliance, Public Records Act, etc.);
- Those legal obligations relating to third party rights in, or over, the data to be stored (e.g. copyright, data protection); and
- 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).

Archives and Copyright: Risk and Reform, CREATe Working Paper No.3

http://www.create.ac.uk/wp-content/uploads/2013/04/CREATe-Working-Paper-No-3-v1-1.pdf

pages 6-18 of this 58 page 2013 paper by R. Deazley and V. Stobo, cover Copyright and the Archive sector within the UK.

A Layman's Guide to the KEEP Legal Studies

http://www.keepproject.eu/ezpub2/index.php?/eng/content/download/20703/103715/file/D2.6_laymansguidelegalstudies_final.pdf

This 2011 paper by D. Anderson of the KEEP (Keeping Emulation Environments Portable) Project (University of Portsmouth, National Library of the Netherlands) discusses the complicated and often contradictory legislative landscape for digital preservation activity in the European Union. Different nation states have their own. Over and above national law, there is the European Community framework, which is not uniformly or completely implemented across the whole of the EU. There is also non-EU legislation, and international treaties and obligations such as the Paris Convention for the Protection of Industrial Property (1883), and the Berne Convention for the Protection of Literary and Artistic Works (1886). The KEEP legal studies threw up two important issues: making copies of digital materials, and making these copies available to users. (39 pages).

Legalities Life Cycle Management

http://timbusproject.net/portal/domain-tools/72-portal/domain-tools/334-lehalities-lifecycle-management-tool

A tool developed by the TIMBUS project which looks at digital preservation of business processes. The areas covered are IPR, IT contracting, Data Protection and other statutory requirements.

Mass Digitization of Cultural Heritage: Can Copyright Obstacles Be Overcome?

http://livestream.com/unc-sils/iPres-Pamela-Samuelson/videos

Keynote presentation from iPRES 2015 by Pamela Samuelson, professor of Law and Information at the University of California, Berkeley. Samuelson has published extensively on IPR and Cyberlaw. In this presentation she considers the role of "fair use" in approaching the challenge that Copyright pose. Samuelson speaks from a US legal perspective but many considerations are also applicable in the UK context. (2015) 56 minutes

iPresKeynote

http://livestream.com/unc-sils/iPres-Pamela-Samuelson

References

Charlesworth, A.J., 2012. Intellectual Property Rights for Digital Preservation, *DPC Technology Watch Report 12-02*. Available: http://dx.doi.org/10.7207/twr12-02

Intellectual Property Office, 2014. *Exceptions to Copyright: Libraries, Archives and Museums*. Available: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/375956/Libraries_Archives_and_Museums.pdf

Koops, B., 2011. *Forgetting Footprints, Shunning Shadows. A Critical Analysis of the "Right To Be Forgotten" In Big Data Practice*. SCRIPTed, 8:3, 229-256. Available:http://script-ed.org/wp-content/uploads/2011/12/koops.pdf

Milligan, I., 2015. *Web Archive Legal Deposit: A Double-Edged Sword*. Available: http://ianmilligan.ca/2015/07/14/web-archive-legal-deposit-a-double-edged-sword/

Risk and Change Management

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Introduction

Digital preservation is not simply about risks. It also creates opportunities and by protecting digital materials it means that new or extended value can be derived from them. It can be easy to become overwhelmed with risks, so It is worth being explicit early in the process about what opportunities are being protected or created. There are many things that put your digital resources at risk including changes to your organisation or technology. If not managed, these risks will have a significant impact on your ability to carry out your digital preservation activities, wider business functions, or comply with legislation.

To manage digital preservation, you must understand your organisation's specific issues and risks. You can do this by undertaking a risk and opportunities assessment. The assessment will highlight specific risks to the continuity of your digital resources, and opportunities that can be realised from mitigating these risks.

Risk management

Experience shows that the risks facing digital resources are subtle and varied. They include, but are not limited to the following:

- Merger, closure, or transfer of functions between organisations.
- Changes in strategic direction or funding and the functions supported by an organisation.
- Major changes in individual leaders or experts.
- Breakdown of resource discovery data resulting in difficulty retrieving data.
- Loss of copyright or other legal information resulting in uncertainty over rights and obligations.

- Outsourcing with no consideration of future preservation needs.
- File format obsolescence meaning that it is expensive or impossible to process data.
- Media obsolescence making it expensive or impossible to recover data.
- Media degradation meaning that data is damaged or changed.
- Loss of contextual information resulting in loss of meaning.
- Loss of provenance information or fixity about a document resulting in loss of authenticity.
- Breakdown of version control making it hard to identify authoritative instances of a document.
- Human error leading to accidental deletion.
- The degree of use. A dark archive is more at risk than one than is heavily used. If digital material is accessed infrequently the impact of failure is less immediately apparent.
- Natural Disasters affecting buildings or infrastructure.

Data loss is likely to have a variety of real world consequences depending on context. In the context of a court case, for example, the authenticity of a document could become a significant legal issue; whereas for highly structured research data the chain of custody may matter less than access to explanatory context that enables the reproducibility of an experiment. In many contexts it may be technically possible to recover digital collections but where an organisation simply doesn't have the wherewithal or skills necessary to restore a data set, then practical obsolescence and data loss can result. This is likely to become more of a reality as the number and complexity of digital collections expand.

The risks to digital content usually matter because of their consequences in the real world. Again this depends on the context but the following can occur:

- Loss of reputation.
- Inadequate resources for a critical task.
- Inability to support users in their activities.
- Failure to discharge legal or regulatory function.
- Inability to exploit and reuse data.
- Loss of identity and corporate memory.
- Cost of recreation and recovery.

Risks are typically prioritised by calculating a 'risk score' based on likelihood, impact and imminence: an imminent risk with a strong probability and a large negative impact needs prompt action. Depending on the nature of the risk this might include taking steps to reduce the likelihood of a risk emerging, reducing the impact if a risk does occur, or buying time for mitigation steps to be implemented.

Risk assessment is an ongoing process that can be developed and expanded through time. It can help bring together different stakeholders and, because risk management is understood by senior management it can also help to make the case for investment. Even an elementary risk assessment will highlight priorities for anyone getting started in digital preservation.

Finally it is worth noting that digital preservation is distinctive in being long-term and most risk methodologies are typically focussed on the short-term. For digital preservation, you need to be aware that over the long term improbable events will become more likely and special attention should be paid to those with significant consequences.

Business continuity planning

Rationale

'Interested parties and stakeholders require that organizations proactively prepare for potential incidents and disruptions in order to avoid suspension of critical operations and services, or if operations and services are disrupted, that they resume operations and services as rapidly as required by those who depend on them.' (ISO/PAS 22399:2007).

Business Continuity planning and practice is well-established within the IT profession and is not dealt with in detail in the Handbook. However it is an important component of ensuring bit preservation and makes a significant contribution to digital preservation through this. There is a series of webinars on business continuity and digital preservation from the TIMBUS project (see Resources).

The development and use of a business continuity plan based on sound principles, endorsed by senior management, and activated by trained staff will greatly reduce the likelihood and severity of impact of disasters and incidents.

One model is the plan developed by the Data Archive, and described in the DPC Case note on Business Continuity. Organisations may also wish to consider use of cloud services (see Cloud services) as part of their planning.

Requirements

- Develop a business continuity plan.
- Ensure all relevant staff are trained in business continuity procedures.
- Create copies of data resources at the time of their transfer to the institution.
- Store copies on industry standard or other approved contemporary media.
- Store copies on and off site. Off-site copies should be stored at a safe distance from on-site copies to ensure they are unaffected by any natural or man-made disaster affecting the on-site copies.
- Consider data and skills as assets and compile registers of them.
- Ensure roles and responsibilities are identified and maintained.

Resources

ISO/PAS 22399:2007 Societal security - Guideline for incident preparedness and operational continuity management

http://www.iso.org/iso/catalogue_detail?csnumber=50295

This standard provides general guidance for any organization to develop its own specific performance criteria for incident preparedness and operational continuity, and design an appropriate management system.

ISO/IEC 27001:2013 Information technology -- Security techniques -- Information security management systems -- Requirements

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

This standard specifies the requirements for establishing, implementing, maintaining and continually improving an information security management system. The requirements are generic and are intended to be applicable to all organizations.

Disaster Preparedness for Digital Content

http://dpworkshop.org/workshops/management-tools/disaster-preparedness

A Digital Preservation Management workshop webpage that links a set of 4 suggested documents (disaster plan policy, communications plan, training plan, roles and responsibilities). Cumulatively they provide comprehensive documentation and are updated to reflect current practice for disaster preparedness.

National Archives Risk assessment tools

http://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/risk-assessment/

The National Archives provide two excel format self-assessment tools that link to its digital continuity guidance and framework of solutions and services.

The Self-assessment tool (0.4 Mb) divides the risk assessment into three sections: Understanding digital continuity and roles and responsibilities; Information requirements and technical dependencies, and Management

The Information asset risk assessment tool (0.26 Mb) helps you identify risks to the continuity of any specific digital information asset and identifies where continuity has already been lost. It makes recommendations on maintaining or restoring continuity to help you develop a digital continuity action plan.

DRAMBORA (Digital Repository Audit Method Based on Risk Assessment) Toolkit

http://www.repositoryaudit.eu

This is an online toolkit for a digital repository audit. The toolkit guides users through the audit process, from defining the purpose and scope of the audit to identifying and addressing risks to the repository. DRAMBORA provides a list of over 80 examples of potential risks to digital repositories, framed in terms of possible consequences.

SPOT

http://www.dlib.org/dlib/september12/vermaaten/09vermaaten.html

The SPOT (Simple Property-Oriented Threat) provides a simple model for risk assessment, focused on safeguarding against threats to six properties of digital objects fundamental to their preservation: availability, identity, persistence, renderability, understandability, and authenticity. The model discusses threats in terms of their potential impacts on these properties, providing several example outcomes for each. The article describing the model also included a useful comparison of other digital preservation threat models.

Managing digital continuity guidance from The National Archives

http://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/

Includes a helpful risk assessment with many correlations to risk management strategies for Business Continuity Planning.

Assess and manage risks to digital continuity

http://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/step-by-step-guidance/step-3/

The National Archives have built a self-assessment tool for the wider public sector that links to its digital continuity guidance and framework of solutions and services.

Assess risks to digital continuity factsheet

http://www.nationalarchives.gov.uk/documents/information-management/assess-dc-risks-factsheet.pdf

(2 pages)

Risk assessment handbook

http://www.nationalarchives.gov.uk/documents/information-management/Risk-Assessment-Handbook.pdf

(35 pages)

The Atlas of Digital Damages

https://www.flickr.com/groups/2121762@N23/

This is a staging area for collecting visual examples of digital preservation challenges, failed renderings, encoding damage, corrupt data, and visual evidence documenting #FAILs of any stripe. You can contribute just an image, tell the story behind the image, or share the original file (or set of files), so that tool developers can learn from digital damage and test out their code with it.

TIMBUS project: Business Continuity Management 1 - Intro, Life Cycle, Planning, Scope

https://www.youtube.com/watch?v=25EhtuE3XkE

1 of 4 Business Continuity Management and the Digital Preservation of Processes webinars from the EU-funded Timbus project. This introduction is probably the most accessible for novices (released 2013. 13 mins).

Case studies

DPC case note: Business continuity procedures – UK Data Archive, University of Essex

http://www.dpconline.org/advice/case-notes/1562-case-note-business-continuity

The Data Archive is the UK national data centre for the Social Sciences funded by the Economic and Social Research Council (ESRC). The Archive holds certification to ISO 27001, the international standard for information security, which requires information security continuity to be embedded in an organisation's business continuity management systems. The digital storage system at the Data Archive is based, for security purposes, on segregated and distributed storage and access. Business continuity at the Data Archive is based around the resilience provided by creating multiple copies of the data and specified recovery procedures, alongside pre-emptive failure prevention. Each file from any dataset has at minimum three copies. The Archive also creates a read only archival copy of each study and any update as it is made available on the system.

References

ISO, 2007. *ISO/PAS 22399:2007. Societal security - Guideline for incident preparedness and operational continuity management.* Available: http://www.iso.org/iso/catalogue_detail?csnumber=50295

Staff Training and Development

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Introduction

A well-skilled and effective workforce can be an organisation's greatest asset, yet due care and attention is not always given to providing adequate training and development, and encouraging its uptake. Additionally, developing and maintaining a digital preservation programme can seem daunting in many ways and, in particular, this is often due to a perceived staff skills gap. This may be because the work environment is characterised by:

- Rapid and ongoing change.
- Blurring of boundaries within and between institutions.
- Uncertainty in terms of the ability to confidently predict future trends and requirements.
- Unclear and/or changing roles and responsibilities.
- Increased emphasis on collaboration and teamwork.
- Increased emphasis on accountability.

Carefully designed staff training and continuous professional development (CPD) activities can play a key role in successfully making the transition from the traditional model of libraries and archives to the digital or hybrid model. Intelligent training and development can do much to boost confidence and ability in staff members, and minimise anxiety about the changing nature of work in preservation-performing institutions. A thoughtful approach to training and development (as opposed to just "sending people on courses") is likely to make a significant difference by:

- Helping staff to exploit technology effectively and improve the overall quality of service.
- Enhancing the individual level of job satisfaction and commitment, leading to improved staff retention.
- Improving the strategic outlook for the organisation as a whole.

Organisations should take a strategic approach to training and development, considering carefully the skills that are required, as well as new and developing roles and responsibilities. The issue should

be clearly addressed in all relevant digital preservation policy, strategy and planning, and budget for advocacy and skills development activities should be an integral part of planning for digital preservation work.

A Broad Range of Skills

Successful digital preservation work requires a broad range of skills, from those specific to the area such as knowledge of metadata standards and audit frameworks, to more general skills such a project planning and risk management. Therefore, ensuring all staff members have adequate digital preservation-specific skills for their part of the process is only one aspect of the preparation required for equipping them to maximise the potential of digital technology. It is highly unlikely that one individual will ever possess all of the skills required to undertake the full range of digital preservation activities, so collaboration will remain key to success. Skilful training can enhance individual skills and competences but can also enhance understanding of the other skills and competences required for a successful collaborative project.

A number of different initiatives have endeavoured to clarify the skills and competencies required for digital preservation work and potential roles involved for staff at different levels of seniority:

DPOE

The Library of Congress's Digital Preservation Outreach and Education programme (DPOE) has defined three levels of staff roles (or career stages) within their model for digital preservation training. These are:

- **Executive** - those in senior institutional management roles.
- **Managerial** - those managing digital preservation programmes and service.
- **Practical** - practitioners working hands-on with digital materials and preservation solutions.

DigCurV

The DigCurV project adapted the DPOE's three level model for their work in defining the core competencies required for digital preservation work. The DigCurV project examined a number of issues relating to digital curation and preservation training, skills and development, producing a variety of useful resources including a database of available training opportunities and a curriculum framework. Describing the core competencies required at each of the three levels in the DPOE model through a set of 'lenses', the DigCurV curriculum framework provides an excellent resource for those looking to identify the full range of skills and competencies required for digital curation and preservation. Specifically, the DigCurV curriculum framework can help users to describe and compare training courses, to develop new training resources and to map the skills and knowledge of an individual or team to identify any existing skills gaps.

Each lens is split into four sections covering

- Knowledge and Intellectual Abilities
- Personal Qualities
- Professional Conduct
- Management and Quality Assurance

Each then contain further sub-sections that list general statements about individual competencies. The statements are designed to be generic so have a broad applicability, although specific examples of particular standards or tools relating to the competencies are available via the version on the DigCurV website.

DigCCurr

The DigCCurr (Preserving Access to Our Digital Future: Building an International Digital Curation Curriculum) project has produced a 6-dimensional matrix for identifying and organizing the material to be covered in a digital curation curriculum. This Matrix of Digital Curation Knowledge and Competencies is an alternative approach that may be particularly useful for smaller organisations.

Roles and Responsibilities for Training and Development

Roles and responsibilities need to be clearly defined. The success of training and development programmes will be affected by the degree to which various roles and responsibilities mesh. It is essential that each of the stakeholders in the process fully appreciate their roles and actively participate in the process. Listed below is a guide to the various responsibilities that may be required of different stakeholders to ensure the creation and deployment of a successful and comprehensive training and development programme.

Stakeholder roles and responsibilities
Roles and Responsibilities of the Institution • Developing an Information Strategy which integrates IT training with the overall mission of the institution. • Identifying, in consultation with key staff, a skills audit, to determine what specific competencies are required to meet organisational objectives, including horizon-scanning for new and emerging skills, activities and responsibilities. • Establishing a balance between recruiting specific skills and effectively developing existing talent. • Providing adequate resources for training and development. • Ensuring staff have access to appropriate equipment. • Ensuring access to practical "hands on" training and practice. • Encouraging networking between colleagues in other institutions. • Considering strategies such as short-term secondment to an institution which may have more experience in a specific area. • Involving staff in designing training and development programmes. • Facilitating effective multidisciplinary communication. • Taking a broad view of what constitutes training and development (i.e. combination of formal courses, both generic and tailor-made, informal training within the organisation, skills transfer within the organisation, networking etc.).
Roles and Responsibilities of Professional Associations

- Responsiveness to current training and development needs.
- Ability to work with institutions to develop training packages to meet their needs.

Roles and Responsibilities of the Individual

- Ability to tolerate frequent change.
- Ability to be flexible.
- Ability to work in teams.
- Ability to communicate (including listening) effectively across staff groups and upwards / downwards within the organisation.
- Ability actively to pursue personal professional development through a range of mechanisms.
- Ability to share skills and expertise.
- Ability to learn new skills.
- Ability to apply new skills.

Undertaking a Skills Audit

A useful starting point for any organisation is to conduct a skills audit tailored to the needs of the specific institution. The process will help identify any skills gaps that exist and allow informed decisions to be made about training and development, as well as potentially highlighting additional roles that may require new staff or new responsibilities (and new job descriptions) for those already in post. Evidence from the skills audit can then be used to build a business case for any additional resources that may be required. In addition to being an excellent starting point for improving staff development it may also be useful to incorporate elements of the process into regular staff professional development and review processes.

The DigCurV curriculum framework or the Matrix of Digital Curation Knowledge and Competencies can provide a useful tool when carrying out a skills audit, in this case as a resource for benchmarking. It will be necessary to tailor the audit to the staff development practices and processes of individual organisations but the following steps may be considered:

1. Identify all roles within the organisation with digital preservation responsibilities. Examining workflows can help in this process and mapping these to models such as the OAIS Reference Model or the Digital Curation Centre Curation Lifecycle Model.
2. Map roles to the relevant lenses of the DigCurV framework.
3. Work with role holders to map skills to the relevant lens. This can be done variety of ways including self-assessment and as a group activity. It may also be useful to mark on a scale.
4. Analyse results to identify gaps, training requirements and additional roles required.

Training and Development Options

A lack of established training and development opportunities was previously a considerable barrier to those wishing to learn more about digital preservation. While those at more advanced levels in

their development may still struggle to find appropriate opportunities, there are now a number of established courses available to those at a beginner and intermediate level from short courses to full degree programmes including a variety of training opportunities addressing specific specialist areas of interest. A greater barrier is now the time and expense involved in attending face-to-face training, but increasingly more online and distance learning options are being made available so this impediment will also decrease.

Digital preservation courses have also previously suffered from criticism relating to an emphasis on theory rather than practice. This too is changing with more practical exercises and tool demos being incorporated into training. Digital preservation also remains a discipline where as much, if not more, can be learnt by doing, so peer to peer learning and a willingness to just get your hands dirty can often produce the best results. Information sharing and short staff exchanges with similar organisations can provide a particularly effect method for staff development and learning.

Resources

APARSEN Survey for the Assessment of Training Material/Assessment of Digital Curation Requirements

http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2012/12/APARSEN-REP-D43_1-01-4_1.pdf

This report presents the findings of research undertaken to in order to set the objectives and strategies for the development of training courses for digital preservation practitioners within the Network of Excellence. The aim of the research was to draw together a comprehensive picture of the digital preservation training that was then available and to identify the training needs of practitioners working within the field. (2012, 109 pages).

2014 DPOE Training Needs Assessment Survey

http://www.digitalpreservation.gov/education/2014_Survey_Report-Final.pdf

An analysis of the state of digital preservation practice and the capacity to preserve digital content within organisations in the United States with the aim of establishing training gaps and needs. (13 pages).

DigCurV, A Curriculum Framework for Digital Curation

http://www.digcurv.gla.ac.uk/

The DigCurV Curriculum Framework offers a means to identify, evaluate, and plan training to meet the skill requirements of staff engaged in digital curation. The DigCurV team undertook multi-national research in the Cultural Heritage sector to understand the skills used by those working in

digital curation, and those sought by employers in this sector. The framework defines separate skills lenses to match the specific needs of three distinct audiences; Executives, Managers, and Practitioners.

- The skills defined under the **Executive Lens** enable a digital curation professional to maintain a strategic view .
- The skills defined under the **Manager Lens** enable a professional to plan and monitor execution of digital curation projects, to recruit and support project teams, and to liaise with a range of internal and external contacts within the cultural heritage sector.
- The skills defined under the **Practitioner Lens** enable a professional to plan and execute a variety of technical tasks, both individually and as part of a multi-disciplinary team.

Matrix of Digital Curation Knowledge and Competencies

http://ils.unc.edu/digccurr/digccurr-matrix.html

The DigCCurr (Preserving Access to Our Digital Future: Building an International Digital Curation Curriculum) project has produced a 6-dimensional matrix for identifying and organizing the material to be covered in a digital curation curriculum.

Digital Preservation Outreach and Education

http://www.digitalpreservation.gov/education/curriculum.html

The Library of Congress's 'baseline' digital preservation training programme for archives and collections management staff. The full course is delivered to archives and other digital preservation professionals in a 'train the trainer' approach, in order to support further dissemination to colleagues. The overview videos are available online.

DPC Training

http://www.dpconline.org/training

A key role for the DPC is to empower and develop its members' workforces. The DPC addresses this issue by facilitating training and support activities and creating practitioner-focused material and events throughout each year. These include The DPC Leadership Programme, The Digital Preservation Roadshow, and The Member Briefing Days and Invitational Events.

Digital Preservation Training Programme (DPTP)

http://dptp.org

A range of UK based digital preservation training courses. Scheduled DPTP courses run over 2 days or 3 days and take place regularly throughout the year.

Digital Preservation Management: Implementing Short-Term Strategies for Long-Term Solutions

http://www.dpworkshop.org/

An excellent free online tutorial that introduces you to the basic tenets of digital preservation. It is particularly geared toward librarians, archivists, curators, managers, and technical specialists. It includes definitions, key concepts, practical advice, exercises, and up-to-date references. The tutorial is available in English, French, and Italian.

UK University post-graduate degree courses

http://www.dpconline.org/training/relevantpostgrads

The DPC maintains a list that will be helpful to anyone looking at post- graduate degrees with a focus on digital preservation. It includes University on campus and distance learning options. Some universities also offer individual credit bearing modules in relevant digital preservation topics.

Education and Training in Audio-visual Archiving and Preservation

http://www.arsc-audio.org/etresources.html

Training opportunities in Australia, Europe and the USA for those working with sound and moving image material.

Connecting to Collections: Caring for Audio-visual Material

http://www.connectingtocollections.org/av/

Self-paced course including recorded webinars, hand-outs, slideshows and suggested further reading for the individual student working with audio-visual material. It covers basic principles, a history of formats and their preservation challenges, format identification, access issues and an overview of existing models and standards. It is written in English by a team of US-based archivists, conservators and digital preservation experts.

How the DPC makes a difference to your staff

https://vimeo.com/45433968

Short interviews with 5 candidates who were sponsored by the Digital Preservation Coalition to attend the Digital Futures Academy in London in March 2012. They reflect on their experience and how joining the DPC has benefitted their institutions. (2 mins 40 secs)

Standards and Best Practice

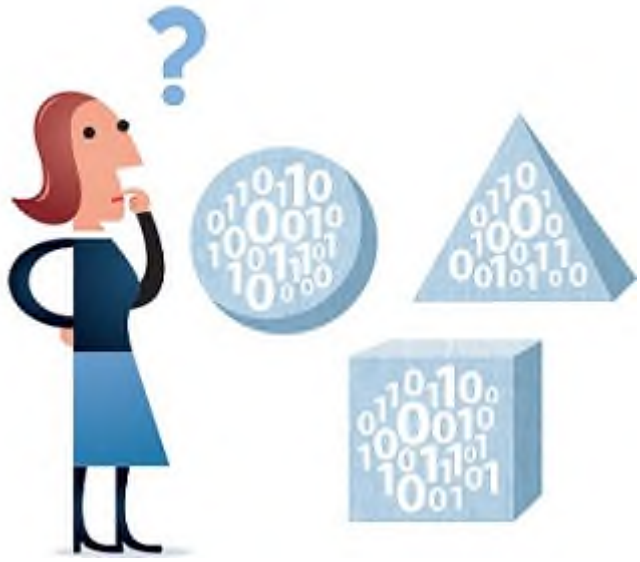


Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Introduction

The use and development of reliable standards has long been a cornerstone of the information industry. They facilitate the access, discovery and sharing of digital resources, as well as their long-term preservation. There are both generic standards applicable to all sectors that can support digital preservation, and industry-specific standards that may need to be adhered to. Using standards that are relevant to the digital institutional environment helps with organisational compliance and interoperability between diverse systems within and beyond the sector. Adherence to standards also enables organisations to be audited and certified.

Operational standards

There are a number of standards which can help with the development of an operational model for digital preservation.

Taking custodial control of digital materials requires a set of procedures to govern their transfer into a digital preservation environment. This can include identifying and quantifying the materials to be transferred, assessing the costs of preserving them and identifying the requirements for future authentication and confidentiality. ISO 20652: Space Data and Information Transfer Systems - Producer-Archive Interface - Methodology Abstract Standard (ISO, 2006) is an international standard that provides a methodological framework for developing procedures for the formal transfer of digital materials from the creator into the digital preservation environment. Objectives, actions and the expected results are identified for four phases - initial negotiations with the creator (Preliminary Phase), defining requirements (Formal Definition Phase), the transfer of digital materials to the digital preservation environment (Transfer Phase) and ensuring the digital materials and their accompanying metadata conform to what was agreed (Validation Phase).

ISO 14721:2012 Space Data and Information Transfer Systems - Open Archival Information System - Reference Model (OAIS) (ISO, 2012b) provides a systematic framework for understanding and implementing the archival concepts needed for long-term digital information preservation and access, and for describing and comparing architectures and operations of existing and future archives. It describes roles, processes and methods for long-term preservation. Developed by the

Consultative Committee for Space Data Systems (CCSDS) OAIS was first published in 1999 and has had an influence upon many digital preservation developments since the early 2000s. A useful introductory guide to the standard is available as a DPC Technology Watch Report (Lavoie, 2014).

An OAIS is 'an archive, consisting of an organization of people and systems that has accepted the responsibility to preserve information and make it available for a defined 'Designated Community'. An 'OAIS archive' could be distinguished from other uses of the term 'archive' by the way that it accepts and responds to a series of specific responsibilities. OAIS defines these responsibilities as:

- Negotiate for and accept appropriate information from information producers;
- Obtain sufficient control of the information in order to meet long-term preservation objectives;
- Determine the scope of the archive's user community;
- Ensure that the preserved information is independently understandable to the user community, in the sense that the information can be understood by users without the assistance of the information producer;
- Follow documented policies and procedures to ensure the information is preserved against all reasonable contingencies, and that there are no ad hoc deletions.
- Make the preserved information available to the user community, and enable dissemination of authenticated copies of the preserved information in its original form, or in a form traceable to the original. (Lavoie, 2014)

OAIS also defines the information model that needs to be adopted. This includes not only the digital material but also any metadata used to describe or manage the material and any other supporting information called Representation Information.

The OAIS functional model is widely used to establish workflows and technical implementations. It defines a broad range of digital preservation functions including ingest, access, archival storage, preservation planning, data management and administration. These provide a common set of concepts and definitions that can assist discussion across sectors and professional groups and facilitate the specification of archives and digital preservation systems.

OAIS provides a high level framework and a useful shared language for digital preservation but for many years the concept of 'OAIS conformance/compliance' remained hard to pin down. Though the term was frequently used in the years immediately following the publication of the standard, it relied on the ability to measure up to just six mandatory but high level responsibilities. A more detailed discussion about 'OAIS compliance' can be found in the Technology Watch Report.

ISO/TR 18492:2005 Long-term preservation of electronic document-based information (ISO/TR, 2005) provides a practical methodology for the continued preservation and retrieval of authentic electronic document-based information, which includes technology-neutral guidance on media renewal, migration, quality, security and environmental control. The guidance is developed to ensure authenticity of records beyond the lifetime of original information keeping systems.

ISO 15489:2001 Information and documentation -- Records management (ISO, 2001) can also be a useful standard for defining the roles, processes and methods for a digital preservation implementation where the focus is the long-term management of records. This standard outlines a

framework of best practice for managing business records to ensure that they are curated and documented throughout their lifecycle while remaining authoritative and accessible.

ISO 16175:2011 Principles and functional requirements for records in electronic office environments (ISO, 2011) relates to electronic document and records management systems as well as enterprise content management systems. While it does not include specific requirements for digital preservation, it does acknowledge the need to maintain records over time and that format obsolescence issues need to be considered in the specification of these electronic systems.

There are international standards that are generic to good business management that may also be relevant in the digital preservation domain.

- Certification against ISO 9001 Quality management systems (ISO, 2015) demonstrates an organisation's ability to provide and improve consistent products and services.
- Certification against ISO/IEC 27001 Information technology -- Security techniques -- Information security management systems (ISO/IEC, 2013) demonstrates that digital materials are securely managed ensuring their authenticity, reliability and usability.
- ISO/IEC 15408 The Common Criteria for Information Technology Security Evaluation (ISO/IEC, 2009) provides a standardised framework for specifying functional and assurance requirements for IT security and a rigorous evaluation of these.

There are a number of routes through which a digital preservation implementation can be certified. These range from light touch peer review certification methods such as the Data Seal of Approval, through the more extensive internal methods of DIN 31644 Information and documentation - Criteria for trustworthy digital archives (DIN, 2012), to the comprehensive international standard ISO 16363:2012 Audit and certification of trustworthy digital repositories (ISO, 2012a) (see Audit and certification).

Technical standards

There are specific advantages to using standards for the technical aspects of a digital preservation programme, primarily in relation to metadata and file formats.

In conjunction with relevant descriptive metadata standards, PREMIS and METS are de facto standards which will enhance a digital preservation programme. PREMIS (PREservation Metadata: Implementation Strategies) is a standard hosted by the Library of Congress and first published in 2005. The data dictionary and supporting tools have been specifically developed to support the preservation of digital material. METS (Metadata Encoding and Transmission Standard) is an XML encoding standard which enables digital materials to be packaged with archival information (see Metadata and documentation).

There are also standards relating to file formats. Choosing file formats that are non-proprietary and based on open format standards gives an organisation a good basis for a digital preservation programme. ISO/IEC 26300-1:2015 Open Document Format for Office Applications (ISO/IEC, 2015) provides an XML schema for the preservation of widely used documents such as text documents, spreadsheets, presentations. ISO 19005 Electronic document file format for long-term preservation (ISO, 2005) prescribes elements of valid PDF/A which ensures that they are self-contained and display consistently across different devices. Aspects of **JPEG-2000** and **TIFF** are also covered by ISO standards. (see File formats and standards).

Barriers to using standards

A standards based approach to digital preservation is important, but there are also factors which inhibit their use as a digital preservation strategy:

- The pace of change is so rapid that standards which have reached the stage of being formally endorsed - a process which usually takes years - will inevitably lag behind developments and may even be superseded.
- Competitive pressures between suppliers encourage the development of proprietary extensions to, or implementations of standards which can dilute the advantages of consistency and interoperability for preservation.
- The standards themselves adapt and change to new technological environments, leading to a number of variations of the original standard which may or may not be interoperable in the long-term even if they are backwards compatible in the short-term.
- Standards can be intimidating to read and resource intensive to implement.
- In such a changeable and highly distributed environment, it is impossible to be completely prescriptive.

These factors mean that standards will need to be seen as part of a suite of preservation strategies rather than the key strategy itself. The digital environment is not inclined to be constrained by rigid rules and a digital preservation programme can often be a blend of standards and best practice that is sufficiently flexible and adapted to suit the needs of the organisation, its circumstances and the digital materials being managed.

Standards, best practice and good practice

In recent years best practice guidance and case studies have been published by national archives, national libraries and other cultural organisations. Digital preservation is also a topic well discussed on blogs and social media which can often provide real time information in relation to theory and practice from around the world. Papers at conferences such as iPRES, the International Digital Curation Conference (IDCC) and the Preservation and Archiving Special Interest Group (PASIG) can be a useful source of up to date thinking from academics and practitioners in digital preservation.

Standards should be understood as a formal description and recognition of what a community of experts might term best practice. Standards, and the best practice from which they derive can be intimidating and there is a risk for those starting in digital preservation that the 'best becomes the enemy of the good'. So in adopting or recommending standards it should always be understood that some action is almost always better than no action. Digital preservation is a messy business which throws up unexpected challenges. So it is almost always the case that a poorly implemented standard is preferable to waiting for perfection.

Sector specific requirements

Specific industries have become active in the development of preservation standards, and particular types of content and use cases have emerged that overlap and extend a number of standards. There is considerable benefit in digital preservation standards being embedded in sector-specific standards since this will greatly assist their adoption, although this may present a challenge to coordination of activities. Three examples are given below:

1. Audio visual materials present a special case for digital preservation (see Moving pictures and sounds). Recommendations for audio recordings and video recordings exist under the

auspices of the International Association of Sound and Audio-visual Archives (such as IASA-TC04, 2009), while a range of industry bodies and content holders including the BBC, RAI, ORF and INA have formed the PrestoCentre to progress research and development of preservation standards in this field. https://www.prestocentre.org/

2. The aerospace industry has particular requirements in product lifecycle management and information exchange which have given rise to a series of industry wide initiatives to standardise approaches to aligning and sharing CAD drawings for engineering. The membership body PROSTEP created the ISO 10303 'Standard for Exchange of Product Model Data' which has developed into the LOTAR standard (http://www.lotar-international.org/lotar-standard/overview-on-parts.html). LOTAR is not incompatible with OAIS, but because it fits within a data exchange protocol important to the industry, aerospace engineers are more likely to encounter LOTAR than OAIS

3. .The Storage Network Industry Association has also begun to make progress on the development of a series of standards. A SNIA working group on long-term data retention has responsibility for both physical and logical preservation, and the creation of reference architectures, services and interfaces for preservation. In addition, a working group on Cloud Storage is likely to become particularly influential in relation to preservation. Cloud architectures change how organizations view repositories and how they access services to manage them. For example, it is unclear how one would measure the success of a 'trusted digital repository' that was based in a cloud provider.

Resources

Seeing Standards; A visualisation of the metadata universe

http://www.dlib.indiana.edu/~jenlrile/metadatamap/seeingstandards.pdf

The sheer number of metadata standards in the cultural heritage sector is overwhelming, and their inter-relationships further complicate the situation. This visual map of the metadata landscape is intended to assist planners with the selection and implementation of metadata standards. Each of the 105 standards listed here is evaluated on its strength of application to defined categories in each of four axes: community, domain, function, and purpose. (2010, 1 page).

Dlib Magazine

http://www.dlib.org/dlib.html

Dlib Magazine publishes on a regular basis a wide range of papers and case studies on the practical implementation of digital preservation standards and best practice.

Data Seal of Approval

http://datasealofapproval.org/en/

PREMIS

http://www.loc.gov/standards/premis/

Library of Congress, 2015

The Digital Curation Centre

http://www.dcc.ac.uk/

The Digital Curation Centre makes available research and case studies in relation to the preservation of research data. Iit also publishes recordings of its annual international digital curation conference proceedings.

The Signal

http://blogs.loc.gov/digitalpreservation/

The Signal is a digital preservation blog published by the Library of Congress

IPRES

http://www.ipres-conference.org/

IPRES, the International Conference on Digital Preservation publishes a website and proceedings from their annual event which looks at different themes within the digital preservation landscape,

The Digital Preservation Coalition Wiki

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

The Digital Preservation Coalition Wiki provides a collaborative space for users of OAIS, the British Library's file format assessments as well as other resources.

Digital Preservation Matters

http://preservationmatters.blogspot.co.uk/

The Digital Preservation Matters blog is a personal account of experiences from working with Digital Preservation

References

DIN, 2012. *DIN 31644 Information and documentation - Criteria for trustworthy digital archives*. Available: http://data-archive.ac.uk/curate/trusted-digital-repositories/standards-of-trust?index=3

IASA-TC04, 2009. *Guidelines in the Production and Preservation of Digital Audio Objects: standards, recommended practices, and strategies: 2nd edition,* edited by Kevin Bradley. Available: http://www.iasa-web.org/tc04/publication-information

ISO, 2001. *ISO 15489:2001 Information and documentation -- Records management*. Geneva: International Organization for Standardization. Available: http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=31908

ISO, 2005. *ISO 19005-1:2005. Document management -- Electronic document file format for long-term preservation*. Geneva: International Organization for Standardization. Available: http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=38920

ISO, 2006. *ISO 20652:2006 Space Data and Information Transfer Systems - Producer-Archive Interface - Methodology Abstract Standard*. Geneva: International Organization for Standardization. Available: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=39577

ISO, 2011. *ISO 16175:2011 Principles and functional requirements for records in electronic office environments.* Geneva: International Organization for Standardization. Available: http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=55791

ISO, 2012a. *ISO 16363:2012 Audit and certification of trustworthy digital repositories.* Geneva: International Organization for Standardization. Available: http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=56510

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

ISO, 2015. *ISO 9001:2015 Quality management systems. Geneva: International Organization for Standardization*. Available: http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=62085

ISO/IEC, 2009. *ISO/IEC 15408:2009 The Common Criteria for Information Technology Security Evaluation*. Geneva: International Organization for Standardization. Available: http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=50341

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

ISO/IEC, 2015. *ISO/IEC 26300-1:2015 Information technology -- Open Document Format for Office Applications (OpenDocument) v1.2 -- Part 1: OpenDocument Schema*. Geneva: International Organization for Standardization. Available:http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=66363

ISO/TR, 2005. *ISO/TR 18492:2005 Long-term preservation of electronic document-based information*. Geneva: International Organization for Standardization. Available: http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=38716

Lavoie, B., 2014. The Open Archival Information System (OAIS) Reference Model: Introductory Guide (2nd Edition). *DPC Technology Watch Report 14-02*. Available: http://dx.doi.org/10.7207/twr14-02

Library of Congress, 2015. *METS Metadata Encoding and Transmission Standard*. Available: http://www.loc.gov/standards/mets/

Business Cases, Benefits, Costs, and Impact

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Introduction

Any change in the economic environment may mean that many organisations are challenged to reduce overall expenditure and to maximise efficiencies. At the same time organisations are preserving increasing amounts of digital material. Reuse of models can form a part of the response to this challenge. The long term management - preservation - of digital materials is an expensive and complex activity. It cannot reliably be done without the investment of resources and expenditure.

The challenges for an organisation are to create business models that:

- Can help define benefits and outcomes and convince key decision makers that it is a worthwhile endeavour;
- Support the wider aims and objectives of the parent organisation;
- Provide for the future of their digital materials in an economically sustainable way and helps reconcile the difference between short to mid-term funding commitments/funding cycles, and long-term preservation goals.

Other organisations have already created templates for business cases and models for the calculation of cost and benefit, so reusing some or parts of these models can not only save time but be used as justification for the adoption of particular strategies.

Business cases

The business case is a tool for advocating and ensuring that an investment is justified in terms of the strategic direction of the organisation and the benefits it will deliver. It typically provides context, benefits, costs and a set of options for key decision makers and funders. It can also set out how success will be measured to ensure that promised improvements are delivered.

It is essential that any business model or proposal that is created supports the wider aims and objectives of the parent organisation. It is equally important that key stakeholders, such as budget holders, are consulted and given early sight of the plans and offered the opportunity to comment and provide input. Early exposure of plans can to some extent mitigate situations in which plans might otherwise be rejected outright.

However, presenting a business case for preserving any material at an early stage is no guarantee that it will be accepted. Whilst there is no sure fire template, some or all of the following steps may be useful if a plan is rejected. Within an organisation there may be set procedures and policies regarding the making and presentation of business cases and these should be followed. Early communication of business planning can help identify topics or areas that could present problems when the plan is formally presented.

Identify options and be pragmatic

The point of business planning is to be aspirational and to create services or products that have value and benefit. Not everyone sees the benefits in preservation over the long term where costs are an ongoing issue or where resources are required to be committed for the long term. Business planning is often an exercise in pragmatism. It might be more effective to make a number of smaller more focussed business plans than one single large proposal. Using their knowledge of an organisation the author of a business plan must ensure that any plan is realistic and within the means of the organisation. Strategic planning provides the framework within which business plans are written. Any strategic objective can be achieved in a number of ways, e.g. less money but more time, fewer staff but longer timeframe etc. A pragmatic response offers decision makers a preferred option and why it is preferred and a small range of other alternative options in the business case. It is often helpful to include the "costs/dis-benefits of inaction" as an option against which other actions can be evaluated.

If at first you don't succeed

Work with stakeholders to identify reasons why a business plan was rejected. Talk to those involved in decision making and seek specific feedback. Was the cost component too expensive? Were the plans too ambitious? Is it felt the business case was poorly written or presented? Does the timeframe not fit with organisational plans?

Response: Work with stakeholders to address key concerns. Be clear to address each issue. Explain the reasons why a business plan was presented and what it is aiming to achieve. Focus on benefits, especially those that address the key strategic goals of the parent organisation. Focus on short as well as longer term benefits of the business plan. One approach is to create business plans that are 'SMART', that is Specific, Measurable, Achievable, Realistic and Timely.

When circumstances change

The hard work in business planning is getting to the point where a plan is accepted. However, circumstances can change. If a business plan is not implemented or previously agreed funding withdrawn, the implications can be severe. Again, communication with key stakeholders is essential and can reveal why something may have changed.

Response: Part of business planning involves having a range of options that can be offered in the event of problems arising with funding a preferred option. Having a well-structured business plan from which proposals can be deleted can help in making an alternative case for phased or alternative implementations requiring fewer resources. In such a case a business plan might quickly be re-drafted in more acceptable terms and resources made available. Having a focus on why resources were not made available gives an opportunity for a business case to be re-presented with more emphasis on benefits and positive impact.

Creating business cases

The following steps should be considered when writing and delivering a business case.

Creating a business case	
1. Audit your digital materials and prioritise work required	Audit your digital materials. Analyse the risks and opportunities for your digital materials. Use your analysis to prioritise areas of work and assign owners to them.
2. Is this the right time?	What are you already doing? Is it the right time to do new things on your own? Can you collaborate with others?
3. Institutional analysis	How ready is your institution for change in terms of content and process?
4. Stakeholder analysis and advocacy	Who will be working on and using the digital materials? Who decides on funding? Engage with them using language and terms they will understand.
5. Objectives: scope aims, activities, plan and costs	Map out what are you going to do, who will do what, what will it cost, and when it will happen.
6. Map benefits to organisational strategy	Make sure you express the benefits of your business case in a way that your funders will understand.
7. What else is needed?	Do you need to include a cost-benefits analysis or list of options based on expenditure and outcome?
8. Validate and refine business case	Review and test your business case against best practice and identify what else it needs.
9. Deliver the business case with maximum impact	Do you have a champion to use in the organisation? Remember you may need to deliver it again.
10. Share an edited business case	Remove confidential material and share online so others can benefit from your work

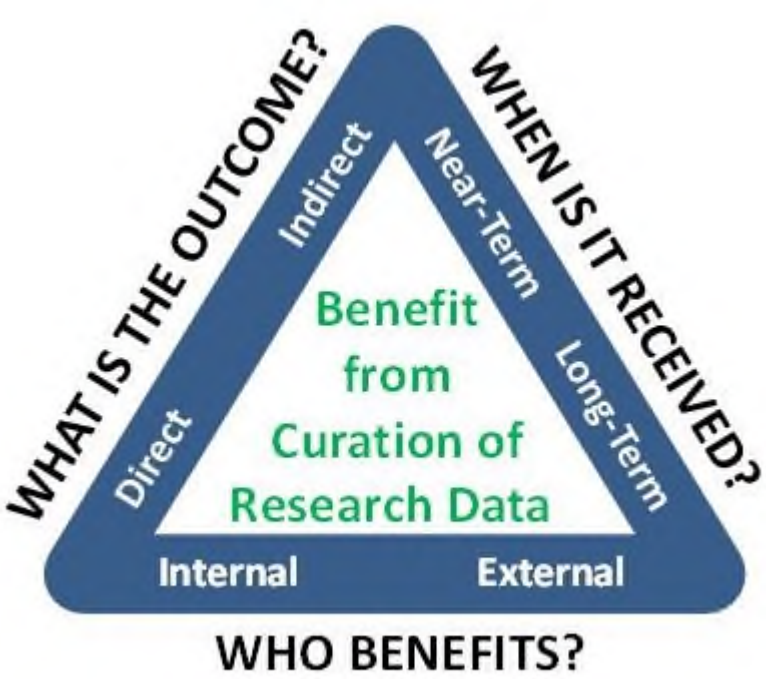
For a generic digital preservation business case template and more information, see the Digital Preservation Business Case Toolkit

Benefits

Benefits are associated with costs and also with risks (see Risk and change management). If risks are mitigated these become a type of benefit. The purpose of the acquisition of any digital material is that it is used. The uses to which digital material is put represents a benefit to those users. If an organisation needs to understand costs associated with digital materials then it must also understand benefits. Benefits can be used to justify costs through the development of business plans.

Measuring benefits is often quite challenging, especially when these benefits do not easily lend themselves to expression in quantitative terms. Often a mixture of approaches will be required to analyse both qualitative and quantitative outcomes and present the differences made. To assist

institutions, the Keeping Research Data Safe project created a KRDS Benefits Framework and a Benefits Analysis Toolkit (KRDS, 2011). These aim to help institutions identify the full scope of benefits from management and preservation of research data and to present them in a succinct way to a range of different stakeholders (e.g. when developing business cases or advocacy). The toolkit is also easily applicable to the benefits of digital preservation to other classes of digital materials.



The KRDS Benefits Framework uses three dimensions to illuminate the benefits investments potentially generate. These dimensions serve as a high-level framework within which thinking about benefits can be organised and then sharpened into more focused value propositions using the Toolkit. It helps you identify what changes you are trying to deliver, what are the outcomes, who benefits, and how long it will take to realise those benefits.

Costs

A business case will normally look at not just the establishment cost for the digital preservation solution, but the all-in cost, including project/program management costs and other activities being undertaken to support implementation such as training and publicity. However digital preservation costs are often the most critical element.

Why understand digital preservation costs?

These are a few reasons why an organisation might want to estimate digital preservation costs:

- Planning and budgeting to build a new repository from scratch, or to extend an existing one.
- Adding a new digital material to your repository and deciding if you can afford it now or over the long-term.
- Providing a platform for comparison with like organisations and an opportunity to adopt efficiencies that have been identified by others.
- Deciding whether to outsource activities or do it in-house.

- Deciding how much to charge for providing a digital preservation service to clients.
- Understanding where resources are being used or under-utilised and areas where additional allocation of resources might be beneficial.

What is lifecycle cost modelling?

A number of research and development projects have sought to model digital preservation costs across the lifecycle from creation and ingest through to preservation and ultimately access. The large number of projects makes understanding this work, finding which results are most applicable to a particular situation, choosing a model, and putting it into practice a significant challenge. The 4C Project surveyed, analysed and assessed this work and provides guidance on getting the most from it:

- Starting out with curation costs - provides an introduction to the concepts.
- Using cost models - describes how to select a cost model appropriate to your organisation.
- Cost Concept Model and Gateway Specification - provides more detail including a guide to develop a model to your own requirements looking at concepts such as 'risk', 'value', 'quality' and 'sustainability'.

Challenges with cost modelling

Cost modelling has been identified as a particularly challenging activity, with a number of difficult aspects, such as:

- Articulating the drivers or aims for costing digital preservation.
- Digital preservation is a moving target, defined by changing technologies and evolving institutional requirements.
- Level of detail: At a high level, modelling becomes less useful as it typically relates to a cross section of different preservation contexts. Modelling at a low level quickly becomes highly complex, making models difficult to develop, maintain and put into use.
- Organisations are reluctant to share costing data, with which models may be developed and validated.
- Even where costing data has been shared, it is often difficult to map between the different costing models it is associated with.
- Separating out digital preservation costs from other business costs is difficult and sometimes meaningless (example: digitising collections in a way that helps ingest into content into the preservation environment –is this digital preservation or digitisation costs?)

For this reason, modelling digital preservation costs across the lifecycle is an activity that should be approached with caution. Cost modelling will always be an approximation and so you need to decide the amount of time you are willing to put in to gain a less approximate answer.

Managing costs

It is possible to manage costs through careful planning. One way is through good process design. The ways in which digital material is created or acquired, managed and disseminated attract costs. Those costs are at the discretion of the organisation and can be managed. The end to end process from acquisition to dissemination must be designed to ensure that all activities are as efficient as possible.

All steps should be designed in such a way as to minimise the need for resources, whilst maximising efficiency. Whilst efficiencies work well at scale, an efficient process doesn't have to be a high volume process. Automation of systematic steps can also save time and deliver effective consistent processes. The initial costs of process design and implementation can be offset by longer term returns.

Impact

Impact is typically the measurement of benefits particularly to the wider public and society undertaken after a business case project has delivered.

For small projects and business cases, impact may be just a simple set of measures such as downloads or number of website requests against which success can be benchmarked easily.

For larger projects and programmes, it may be part of a more thorough evaluation to justify the resources expended. It can include a mixture of quantitative and qualitative measures and will normally be undertaken by external specialists working with staff from the repository. They employ methods from economics and management and information science, for example cost-benefit analysis or contingent valuation, and traditional social science methods such as interviews, surveys and focus groups.

Measurement involves choosing metrics or indicators and requires careful planning and agreement about what to measure and how. Metrics often employ readily countable things such as downloads, or scales metrics that are not truly numeric, such as rating scales or categories of variables. Typically there is a trade-off between what ideally should be measured (e.g. users and use) and proxy measures which are easy to capture and measure (e.g." unique visitors" and web downloads).

Resources

Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information

http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf

The Blue Ribbon Task Force investigated sustainable digital preservation and access from an economic perspective. This final report, identifies problems intrinsic to all preserved digital materials, and proposes actions that stakeholders can take to meet these challenges to sustainability. It developed action agendas that are targeted to major stakeholder groups and to domain-specific preservation strategies. (2010, 116 pages).

The Value and Impact of Data Sharing and Curation: A synthesis of three recent studies of UK research data centres

http://repository.jisc.ac.uk/5568/1/iDF308_-_Digital_Infrastructure_Directions_Report%2C_Jan14_v1-04.pdf

This synthesis summarises and reflects on the combined findings from a series of independent investigations into the value and impact of three well established UK research data centres or services (the Economic and Social Data Service, the Archaeology Data Service, and the British Atmospheric Data Centre). The studies adopted a number of approaches to explore the value and

impacts of research data services and the data sharing and archiving that they have enabled. Data collection involved focused user and depositor surveys, and data centre financial and operational data (e.g. user registrations, dataset deposits and downloads), supplemented by in-depth interviews. Not all impacts can be captured and quantified; therefore they have used these economic approaches with others, such as the KRDS Benefits Framework, to illustrate wider benefits. (2014, 26 pages).

Jisc DigitalMedia infokit on Sustainability and Funding

http://www.jiscdigitalmedia.ac.uk/infokit/digitisation-funding-and-sustaina/digitisation-funding-and-sustaina-home

Provides a starting point for considering the issues necessary to create and build a business model that will support sustainability of digitisation and digital collections.

4C Project Collaboration to Clarify the Costs of Curation

http://4cproject.eu/

The European Union funded 4C project aimed to help organisations across Europe to invest more effectively in digital curation and preservation. A series of reports and resources were produced and are available from its outputs and deliverables page. These include the Digital Curation Sustainability Model , an Evaluation of Cost Models and Needs & Gaps Analysis, a Report on Risk, Benefit, Impact and Value, and a Draft Economic Sustainability Reference Model. The evaluation of costs models report evaluates ten available cost models including, KRDS and LIFE. Another major output was the Curation Costs Exchange (CCEx), a community owned platform which helps organisations of any kind assess the costs of curation practices through comparison and analysis. The CCEx aims to provide real information about costs to help make more informed investments in digital curation. The CCEx was launched in 2014 by 4C and is now maintained and governed by the Digital Preservation Coalition (DPC) with help from nestor and The Netherlands Coalition for Digital Preservation (NCDD).

Digital Preservation Business Case Toolkit

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

This Toolkit provides an in depth guide to writing a business case that is focused on digital preservation activities. It's targeted at practitioners (and their managers) who are working with digital resources and would like to obtain funds to expand their digital preservation activities. The Toolkit is primarily aimed at those seeking further funds from within their organisation, but could also provide useful information for those writing a bid for project funds from an external funding body. It includes a Step by step guide to building a business case and a Template for building a business case. Created by the Jisc funded SPRUCE Project in 2013 the toolkit wiki is hosted by the DPC.

Keeping Research Data Safe (KRDS) Benefits Toolkit

http://www.beagrie.com/krds/

Keeping Research Data Safe (KRDS) is a series of cost/benefit studies, tools and methodologies that focus on the challenges of assessing costs and benefits of curation and preservation of research data. Although focussing on research data, the tools are easily customised to apply to other areas of digital preservation. Available outputs include a KRDS Factsheet, a KRDS User Guide, a KRDS Activity Cost Model, and a KRDS Benefits Analysis Toolkit as well as supplementary materials and reports. The KRDS projects between 2008 and 2011 were funded by Jisc.

20 Cost Questions for Digital Preservation

http://www.metaarchive.org/public/publishing/ma_20costquestions_final.pdf?thumblink

The MetaArchive Cooperative has produced a set of 20 questions to "assist institutions with their comparative analysis of various digital preservation solutions". This work marks a move away from the development of detailed predictive costing models towards a more general approach that seeks to identify and understand key cost drivers rather than the actual costs themselves.

DSHR's Blog

http://blog.dshr.org/search/label/storage%20costs
http://blog.dshr.org/search/label/cloud%20economics

David Rosenthal is a frequent blogger on the topic of storage costs, often considering the impact of the evolution of storage technology on preservation costs and on cloud storage.

A Digital Asset Sustainability and Preservation Cost Bibliography

http://blogs.loc.gov/digitalpreservation/2012/06/a-digital-asset-sustainability-and-preservation-cost-bibliography/

A bibliography that "ranges broadly, from articles on "contingent valuation," "ecosystem valuation" and the general "costs" of knowledge, to those that directly address the cost issues associated with digital preservation and stewardship".

Digital Preservation and Data Curation Costing and Cost Modelling

http://wiki.opf-labs.org/display/CDP/Home

A list of digital preservation cost models and cost modelling initiatives.

The Cost of Inaction Calculator Rationale

https://coi.avpreserve.com/rationale

This is a great information video from AVPreserv on the cost of inaction and the business case rationale for digital preservation. It is focussed on Audio-Visual material but it worth listening to and thinking laterally about the underlying rationale whatever type of digital material you hold. (8mins 41sec)

Case studies

KRDS Benefits Toolkit case studies

There are 4 case studies providing worked examples of completed worksheets from project partners as follows:

Archaeology

http://www.ukoln.ac.uk/events/i2s2-krds/presentations/catherine-hardiman-krds-benefit-framework-2011-07-v2.ppt

The background to this case study is provided in the Archaeology Data Service (ADS) dissemination workshop presentation. Worked examples are available of the ADS Benefits Framework Worksheet (PDF) and the ADS Value-chain and Impact Worksheet (Excel 97-2003).

Health: Population Cohort Studies

http://www.ukoln.ac.uk/events/i2s2-krds/presentations/dipak-kalra-krds-benefits-2011-07.ppt

The background to this case study is provided in the Medical Research Council Cohort Studies dissemination workshop presentation. A worked example is available of the Cohort Studies Value-chain and Impact Worksheet (Excel 97-2003).

Research Data Citation: SageCite

http://www.ukoln.ac.uk/events/i2s2-krds/presentations/monica-duke-krds-sagecite-benefits-2011-07.ppt

The background to this case study is provided in the SageCite dissemination workshop presentation. A worked example is available of the SageCite Benefits Framework Worksheet (PDF).

Social Sciences: UK Data Archive (UKDA)

http://www.ukoln.ac.uk/events/i2s2-krds/presentations/matthew-woollard-krds-benefits-2011-07.ppt

The background to this case study is provided in the UKDA dissemination workshop presentation. A worked example is available of the UKDA Benefits Impact Worksheet (PDF).

Digital Preservation Business Case Toolkit

There are four case studies as follows sourced from activities conducted as part of SPRUCE Project Awards:

Bishopsgate library case study

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

A collections audit and business case focused on taking the first steps of digital preservation.

Institute of education case study

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

A review of approach and generation of a business case for digital administrative record keeping.

Northumberland estates case study

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

Assessment of digital repository solutions and an associated business case for digital preservation.

Lovebytes case study

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

A trial of media stabilisation and content preservation along with a business case to move to a production status.

References

Keeping Research Data Safe (KRDS), 2011. *Digital Preservation Benefits Analysis Toolkit*. Available: http://beagrie.com/krds-i2s2.php

Digital Preservation **Handbook**

Organisational Activities

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Who is it for?

Creators and publishers of digital resources, third-party service providers, operational managers (DigCurV Manager Lens) and staff (DigCurV Practitioner Lens) with responsibility for implementing institutional activities of relevance to digital preservation. It is assumed that these will include a) staff from structurally separate parts of the organisation, and b) a wide range of knowledge of digital preservation, from novice to sophisticated; c) both technical and non technical perspectives; d) a wide range of functional activities with a direct or indirect link to digital preservation activities.

Assumed level of knowledge

Wide-ranging, from novice to advanced.

Purpose

- To provide pointers to sources of advice and guidance aimed at encouraging good practice in creating and managing digital materials. The importance of the creator in facilitating digital preservation is stressed throughout the handbook but particularly in Creating digital materials. Good practice in digitisation and other digital materials creation is crucial to the continued viability of digital materials.
- To raise awareness of factors which need to be considered when creating or acquiring digital materials.
- To provide pointers to helpful sources of advice and guidance for both novices and those who have already begun to think through theimplications of digital technology on their operational activities.

Reusing this information

Contents

Creating digital materials

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Introduction

"The first line of defense against loss of valuable digital information rests with the creators, providers and owners of digital information." (Waters and Garrett, 1996)

The Task Force on Archiving of Digital Information articulated one of the earliest acknowledgements of the crucial role of the creator in helping to ensure long-term access to the digital resources they create. This view has been reiterated in many other documents since.

Clearly, most individual creators cannot be expected to take on a long-term commitment to preserving the digital content they create beyond that of their business needs. Every digital resource has a life cycle and different stakeholders and interests within this. However, it is both highly desirable and achievable that a dialogue is established between long-term repositories and creatorswhen issues of long-term preservation are involved. It is often in the creator's interest as well that content created is well-formed, complete, correct, and usable for current and future purposes. Given the crucial role of the creator in undertaking short to medium-term preservation often for a period of decades and at least facilitating medium to long-term preservation, encouraging good practices (and also outreach by repositories), are crucial.

This section will focus solely on encouraging good practices in the creation of digital materials which will assist in their longevity of active use, future management and preservation. You should refer to other relevant sections of the Handbook for related activities and guidance.

Our focus remains the generic implications for digital preservation in the creation process of digitisation (digital surrogates) or that of born-digital materials.

Creating digital surrogates or domain specific types of born-digital files such as electronic records, research data, or personal digital information all have excellent sources of further advice and guidance. Key references are provided in Resources and case studies.

Creating born-digital materials

Digital preservation refers to the series of managed activities necessary to ensure continued access to digital materials for as long as necessary. This includes the activities when creating born-digital materials necessary to meet the ongoing needs of the original creator.

Often many of these actions needed for continuing access for the long-term overlap with those best practices suited to immediate business needs. Indeed many organisations and individuals create digital materials now that they will need to use and manage for many decades. They would probably not consider themselves as doing digital preservation and other terms such as "digital continuity" are frequently used to communicate how these actions affect them when they are not memory organisations such as museums, libraries or archives with a mission to preserve.

It is important for creators to realise if they do not actively work to ensure continuity, their digital materials can easily become unusable. It is about making sure that their information is complete, available and therefore usable for their business needs.

Your information is usable if you can:

- Find it when you need it;
- Open it as you need it;
- Work with it in the way you need to;
- Understand what it is and what it is about;
- Trust that it is what it says it is.

This enables you to operate accountably, legally, effectively and efficiently. It helps you to protect your reputation, make informed decisions, reduce costs, and deliver better services. For further information on first steps in digital preservation see Getting started.

The following table provides guidance on key issues and actions to consider when creating digital materials to ensure their longevity of active use and potential for long-term preservation.

Preserving born-digital materials	
1	**Software and formats** Choose software that is well supported and creates files that can be read by a variety of different programs. See the File formats and standards section of the Handbook for relevant guidance.
2	**File names** Use a short descriptive file name of content and date that provide context and can be easily understood by humans and computers, now and in the future. Do not use spaces or special characters (other than - or _), this will avoid potential mis-interpretation by computer hardware or software. Put date information in the ISO 8601:2004 standard format: YYYY-MM-DD. This provides a consistent method for version tracking. Note separate file date metadata generated by systems can often change automatically with later actions. Use a consistent method for showing the file versions. This can be date as above, supplemented as needed , e.g.by a version number v1, v2, v_final, etc.

3	**Storage and backup** See Principles for using IT storage systems for digital preservation in the Storage section of the Handbook for relevant guidance.
4	**Know your obligations and relevant best practice** See the Legal compliance section of the Handbook for relevant guidance. Many obligations and best practices will be project or sector specific – see the Creating Research Data inset below for an example.
5	**Plan for transitions** Some transitions can be foreseen and planned for others may be unforeseen but can be mitigated by good planning and procedures. See Resources and case studies below and the Preservation planning and Risk and change management sections of the Handbook for relevant guidance.

Creating digital surrogates

The emphasis on digitisation in this section reflects its current importance as increasing numbers of institutions embark on digitising parts of their collections. It is important to reinforce that this Handbook is not considering the potential of digitisation as a preservation reformatting tool. The emphasis is on the preservation of born-digital materials, or the products of digitisation (the digital surrogates themselves), not the preservation of the analogue originals.

One important exception is that for audiovisual materials. Audio and video materials need digitisation for the very survival of their content, owing to the obsolescence of playback equipment and decay and damage of physical items, whether analogue or digital (see Moving pictures and sound).

Many digitisation projects cite enhanced access as the major objective, a perfectly legitimate objective but unless due care and attention is given to how that access can be maintained over time, it may well be short-lived. It is unlikely that all current digitisation initiatives are being undertaken with due regard to the long-term viability of the digital surrogates they are creating, so it is useful to encourage good practice in creating digital materials and to point to existing sources of guidance.

A good portion of what is now being digitized began life as born digital content. It was converted into an analogue format such as print on paper before the need for digital access and to re-digitize was recognised. That cycle needs to shift quickly to simply managing more born digital content.

Preserving digital surrogates: digital preservation considerations	
1	**Assessment of need for digitisation** Has the material already been digitised? If so, is it to an appropriate standard and readily accessible to your audience?
2	**Finding funds for the project**

	What archiving policies exist, both from the funding agency (if externally funded) and the institution with prime responsibility for the project?
3	**Planning the project and assigning resources** Need to set aside recurrent funds for maintenance of the digital copies as well as one-off funds for conversion. --- Ensure all relevant stakeholders are aware of the project (for example, if another part of the organisation or an external agency is expected to maintain the resource, they will need to be included in discussions at this point, if not before) Identify a strategy for carrying forward the assets of the project in a sustainable manner after the project has achieved its deliverables. This strategy might involve ingesting the assets of the project into the collection catalogue of the parent organisation, or designating a partner institution for receipt of these assets.
4	**Selection of materials** Copyright. It will be necessary to ensure permission is given both to digitise the original and to make copies of the digital copy for the purposes of preservation and delivery. For further information, see Legal compliance. --- Condition and completeness of original. Is it capable of being re-scanned at a later date if the digital copy is lost?
5	**Decide how the information content needs to be organised** (for example, searchable text databases and/or document page images) Selection of appropriate file formats and storage for both master/archive copies and derivatives, see File formats and standards, Metadata and documentation and Storage.
6	**Decide digitisation method appropriate to analogue original and goals of the project.** Preparing originals for digitisation Details of the digitisation method need to be documented and attached to the metadata record to enable future management.
7	**Preparing originals for digitisation** The National Archives provides standards and guidance on document preparation for digitisation of records (The National Archives, 2015). Will the originals be retained? Do not to take any action on discarding the originals until it is established that a) the electronic version is legally admissible and/or b) the electronic version is capable of long-term preservation. Deciding whether or not to retain the originals post-digitisation will of course not be an issue for projects digitising valuable treasures within a collection, the main issue then will be whether or not the original is too fragile to be re-scanned at a later date if the digital copy is

	lost. In any of these cases, if the digital copy becomes the primary means of access, it will be subject to the same requirements as born digital material.
8	**Conversion** Documentation of technical characteristics. Compression algorithm (if used); bit depth required; scanning resolution etc. Create backup copies as soon as conversion is undertaken.
9	**Quality assurance checks** Digital surrogate needs to be of an acceptable preservation quality. --- If using third party services, need to ensure documentation clarifies responsibility for quality assurance.
10	**Final indexing and cataloguing** Metadata for resource discovery and for managing and preservation of digital copy.
11	**Loading data into computer systems** Document storage requirements for access and preservation copies (if different). Make backup copies as appropriate.
12	**Implementing archiving and preservation strategies or transferring to a preservation agency** Required standards for formats, storage media, documentation, and transfer procedures. Storage of masters and backup copies. Strategies for media refreshment and changes in technological environment.

Resources

Digitisation at The National Archives

http://nationalarchives.gov.uk/documents/information-management/digitisation-at-the-national-archives.pdf

This document sets out TNA's standards and requirements for the digitisation of analogue records in its collection. It is also recommended to UK government departments who wish to digitise any of their paper records It covers: the whole digitisation process from initial scanning through to delivery of the images for preservation, including The National Archives' scanned image specification; the scanning of records where the resultant images will become the legal public record for permanent preservation; and the scanning of records where the resultant images will become digital surrogates with the original paper records being retained and remaining the legal public record (July 2015, 56 pages).

Koninklijke Bibliotheek/National Library of the Netherlands: Metamorfoze preservation imaging guidelines

http://www.metamorfoze.nl/sites/metamorfoze.nl/files/publicatie_documenten/Metamorfoze_Preservation_Imaging_Guidelines_1.0.pdf

Metamorfoze is the national program of the Netherlands for preserving paper heritage. The guidelines are intended for the digitisation of two-dimensional materials such as manuscripts, archives, books, newspapers and magazines. They may also be applied to photographs, paintings and technical drawings. The Guidelines relate exclusively to the image quality and metadata of the Preservation Master file, from which all outputs intended for print and/or the web can be derived. (2012, 44 pages).

Preparing Collections for Digitisation

This 2010 book by Anna E. Bulow and Jess Ahmon offers practical guidance covering the end-to-end process of digitising collections, and can be used as a 'how-to' reference manual for collection managers who are embarking on a digitisation project or who are managing an existing project. It also covers some of the wider issues such as the use of surrogates for preservation, and the long term sustainability of digital access. (208 pages).

InterPARES 2 Creator Guidelines Making and Maintaining Digital Materials

http://www.interpares.org/ip2/display_file.cfm?doc=ip2(pub)creator_guidelines_booklet.pdf

This booklet provides advice for individuals who create digital materials in the course of their professional and personal activities to help them ensure their preservation (10 pages).

InterPARES 2 Preserver Guidelines Preserving Digital Records: Guidelines for Organisations

http://www.interpares.org/public_documents/ip2(pub)preserver_guidelines_booklet.pdf

This booklet provides advice to any organization responsible for the long-term preservation of digital records (10 pages).

Jisc Digital Media resources

Copyright and still images: Frequently Asked Questions

http://www.jiscdigitalmedia.ac.uk/guide/copyright-and-still-images-frequently-asked-questions

infokit: Digitisation funding and sustainability

http://www.jiscdigitalmedia.ac.uk/infokit/digitisation-funding-and-sustaina/digitisation-funding-and-sustaina-home

infokit: High level digitisation for audiovisual resources

http://www.jiscdigitalmedia.ac.uk/infokit/audiovisual-digitisation/audiovisual-digitisation-home

infokit: Still image digitisation

http://www.jiscdigitalmedia.ac.uk/infokit/digitisation/still-image-digitisation-home

infokit: Digital file formats

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

Federal Agencies Digitization Guidelines Initiative

http://www.digitizationguidelines.gov/

This is a collaborative effort by US federal agencies to define common guidelines, methods, and practices for digitizing historical content. As part of this, two working groups are studying issues specific to two major areas, Still Image and Audio Visual.

Future Proof – Protecting our digital future

http://futureproof.records.nsw.gov.au/

Future Proof is a State Records initiative from the New South Wales State Government in Australia. This website and blog cover products and projects from State Records that are specifically about digital records. Category links provide useful gateways into different resources housed on the site.

The Curation Reference Manual

http://www.dcc.ac.uk/resources/curation-reference-manual

This resource maintained by the Digital Curation Centre contains advice, in-depth information and criticism on current digital curation techniques and best practice. The Manual is an ongoing, community-driven project, which involves members of the DCC community suggesting topics, authoring manual instalments and conducting peer reviews. Each instalment is designed to help data custodians, producers and users better understand the challenges they face and the roles that they play in creating, managing and preserving digital information over time.

UK Data Archive: Create and Manage Data - Ethical / Legal / Overview

http://www.data-archive.ac.uk/create-manage/consent-ethics/legal

Collecting, using and sharing data in research with people requires that ethical and legal obligations are respected. Laws such as the Data Protection Act, Freedom of Information Act and Statistics and Registration Services Act also govern the use of some kinds of data. This guidance offers help on how research data can be shared without breaching ethical or legal responsibilities.

An Elevator Pitch for File Naming Conventions

http://acrl.ala.org/techconnect/post/an-elevator-pitch-for-file-naming-conventions

This Association of College and Research Libraries (ACRL) TechConnect blog post makes the case for adopting a consistent approach when naming digital files or software components, by demonstrating the effects of not doing so. (2013).

Digital Continuity guidance from The National Archives UK

http://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/

Comprehensive guidance on digital continuity from the National Archives. Of particular use are: Understanding Digital Continuity an introduction to the topic (2011, 20 pages); and Managing Digital

Continuity which takes you through a 4 stage process:1) Plan for action 2) Define your digital continuity requirements 3) Assess and manage risks to digital continuity 4) Maintain digital continuity.

NCDCR Digital Preservation Best Practices and Guidelines - Create Digital Files

http://digitalpreservation.ncdcr.gov/

First launched in 2010 by the Digital Information Management Program of the State Library of North Carolina, and the Digital Services Section of the State Archives of North Carolina, this site received a National Digital Stewardship Alliance Innovation Award in 2012. The aim is to provide practical, introductory information about digital preservation, and to direct visitors to approachable "next step" resources.

Digital Preservation Management Tools and Techniques

http://dpworkshop.org/workshops/management-tools

DPM workshop content workflows include a high-level diagram with lower-level diagrams for managing physical content, transitioning through digitisation, and managing born-digital and digitised content. The idea is to provide a common workflow for all content in any context then develop a number of use cases to highlight exceptions for specific kinds of content with different kinds of requirements.

Part 1: Why is File Naming Important?

https://www.youtube.com/watch?v=Hi_A4Ywn4VU

This excellent short video is part one of a four-part tutorial on file naming. It talks about why it's important to choose your file names wisely. Designed for a general audience, it is part of the State Library of North Carolina's "Inform U" series (2012, 3mins 19 secs).

Case studies

DPC case note: ULCC assessing long term access from short term digitization projects

http://www.dpconline.org/component/docman/doc_download/534-casenoteassessingpreservationindigitization.pdf

Digitisation projects are mostly funded over a short term, so how can we take steps to make the outputs of digitisation robust in the long term? This Jisc-funded case study reports work undertaken by the University of London Computer Centre in assessing the long term plans of 16 digitisation projects, providing a basic survey tool to help funders and project managers alike to reflect on their long term preservation plans. November 2010 (4 pages).

The British Library 'Save our Sounds' project

http://www.bl.uk/projects/save-our-sounds

Launched in 2015 Save our Sounds is the British Library's programme to preserve via digitisation the nation's Sound Archive, a collection of over 6.5 million recordings of speech, music, wildlife and the environment, from the 1880s to the present day. The project aims both to ensure that the existing archive is properly preserved, and that there are adequate systems in place for the acquisition of future sound production in the UK.

Digital Curation Centre case studies

http://www.dcc.ac.uk/resources

In 2013 the DCC began a series of case studies to accompany the new DCC guide How to Develop Research Data Management Services. These cover specific components of a Research Data Management service of interest to researchers and data managers.

Society of American Archivists campus case studies

http://www2.archivists.org/publications/epubs/Campus-Case-Studies

Campus Case Studies are reports by American university archivists who have created working solutions. They cover a wide range of topics some of which are specifically focussed on digital preservation and creating digital records. The currency of the case studies varies from 2008 to the present.

Why metadata matters

https://cbaileymsls.wordpress.com/2013/09/29/metadata/

This blog post provides good examples of why poor file-naming and metadata description at creation of a file can hinder subsequent searching, discovery and re-use.

References

The National Archives, 2015. *Digitisation at The National Archives*. Available: http://nationalarchives.gov.uk/documents/information-management/digitisation-at-the-national-archives.pdf

Waters, D and Garrett, J., 1996. *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information commissioned by the Commission on Preservation and Access and the Research Libraries Group*. Washington, DC: Commission on Preservation and Access. Available: http://www.oclc.org/programs/ourwork/past/digpresstudy/final-report.pdf

Acquisition and Appraisal

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Introduction

In a digital environment, decisions taken regarding creation and selection have significant implications for preservation. The link between access and preservation is far more explicit than for paper and other traditional materials, as access to a digital resource can be lost within a relatively brief period of time if active steps are not taken to maintain (i.e. preserve) it from the beginning. As the interactive Decision Tree indicates, if it is neither feasible nor desirable to preserve a digital resource across various changes in technology, then its acquisition may need to be re-evaluated. While many of the same principles from the traditional preservation environment can usefully be applied, policies and procedures will need to be adapted to the digital environment.

In a print environment, the physical dimensions of an archive mean the reasons to select are relatively well understood, while the decision to preserve can be taken quite separately and within a timeframe which may span several decades. In contrast, digital resources can proliferate and appraisal can be a daunting task. Moreover they can become inaccessible relatively quickly, so decisions about selection and preservation may need to be taken simultaneously for digital collections.

While this may mean that greater rigour is required in selecting digital resources than for printed or other analogue material, it will avoid costs which will otherwise occur later as retrospective preservation of digital resources is not recommended.

Accurate documentation is also crucial in the digital environment. This will provide not only essential details for managing the resource over time but also information on context without which there may be little point in preserving the digital object itself even if it is technically feasible to do so. In the accompanying Decision Tree, it is suggested that acquisition be re-evaluated if documentation is inadequate.

In the case of networked digital resources, where providing access to a resource does not necessarily require bringing the resource physically into a collection, the concept of acquisition is quite different

from traditional collections. There are a range of options available to provide access or to build 'virtual collections'. For example, making copies/mirrors for access, providing a hyperlink to a resource, online catalogues and finding aids.

In some cases an institution may be reluctant to take primary preservation responsibility for material if it feels that interest in its preservation is so widely shared that it would constitute an unfair burden on their own institution. This emphasises the need for collaboration between institutions and the need to establish equitable agreements for shared efforts where necessary. A number of services have emerged in recent years, like the Keepers Registry for electronic journals or the DLF/OCLC Registry of Digital Masters which allow institutions to identify preservation intent - their commitment to preserve material that may be of general interest. The accompanying Decision Tree for appraisal and selection is based on the assumption that the resource has not yet been acquired and indicates a number of points at which cost implications will need to be taken into account before the decision to proceed with acquisition. It suggests that, at these points, difficult choices may need to be made about whether the resource justifies the costs or whether it is preferable not to proceed with acquisition.

Development of policy and procedure

Before embarking on the acquisition and ingest of digital collections it may be necessary to establish whether current policies (e.g. collection development) and procedures are still fit for purpose in the digital age (see Institutional policies and strategies) . Depending on the structure and wording of existing documents this review may result in anything from small changes that increase scope to include digital objects through to the creation of new policy documents specifically covering digital collections. Additions or alterations to the policy may include descriptions of the types of objects that will be acquired, in relation to format and/or content, as well as addressing other issues such as intellectual property rights, sensitivity and access considerations. It is essential that any changes are ratified by the relevant management committees within your organisation to ensure support and buy-in.

Appraisal/retention policy

The Decision Tree accompanying this section may be used as a tool to construct or test the selection, or appraisal/retention policy for your organisation.

Appraisal of born digital objects should include a measured assessment of their value to the parent organisation set against the challenges of long-term preservation and providing access. These challenges may include an organisation's ability to read or open a version of the master file, the ability to secure sufficient rights to manage and provide access to current and future versions of the file, or simply staffing and funding resources. Organisations should therefore initially focus on a balance between acquisition of high value digital objects and these longer term curatorial obligations. It should be remembered that organisations can provide access to resources that they have accessioned without placing them in specific preservation or retention workflows. A detailed policy document which clearly identifies the most important digital resources (from either a format or content perspective) can give guidance on appraisal of born digital objects destined for such pathways. For lesser value digital acquisitions, which often come bundled with higher value acquisitions, it may be enough to outline the level of access and preservation an organisation will provide to them. This outline should include an indication of a retention schedule suitable to this type of content. This may mean also including a disposal schedule or de-accessioning policy if appropriate (see Retention and review).

Agreements and Guidance for depositors - file formats, required documentation

Once policy has been established there will be a number of additional supporting documents that will be required to facilitate the acquisition and appraisal process. Alongside the standard procedural documents an organisation may wish to create a suite of standard depositor agreements and licences to aid in the negotiation process. These will be particularly useful in ensuring that the minimum permissions and intellectual property rights required for preservation are granted. Without sufficient licence agreements an organisation may find itself in possession of digital collections that it does not hold the rights to actively preserve or provide access to (see Legal compliance). These may also be complemented by guidance notes for depositors that set out requirements for material to be transferred and accompanying documentation.

Standards for acquisition and transfer

Experience shows that the transfer from a producer to an archive can be tortuous and therefore any tools which can streamline the process are likely to benefit both sides. Two initiatives have attempted to standardise the interface between Producers and Archives into a consistent, well-understood process, cultivating a mutual understanding between producers and archives in regard to their respective roles: Producer-Archive Interface – Methodology Abstract Standard (PAIMAS, ISO 20652:2006); and the Producer–Archive Interface Standard (PAIS, ISO 20104:2015).

PAIMAS provides a standardized description of the interactions between producers and an archive. It segments the transfer process into a number of phases, providing a detailed description of the anticipated outcome of each phase and the actions required to bring about this outcome. The four principle phases - Preliminary phase, Formal Definition Phase, Transfer Phase, Validation Phase - serve as a basis for identifying areas within the Producer-Archive interface that would benefit from more focused standards, recommendations, and best practices, and also provides a foundation for the development of automated processes and software tools to support the information transfer process. PAIMAS implicitly expands the detailed requirements for Ingest and Administration within the OAIS reference model.

PAIS provides a standard method for formally defining the digital information objects to be transferred by an information Producer to an Archive and for effectively packaging these objects in the form of Submission Information Packages. It is intended to support more precise definitions of the digital objects, helping archives process and validate objects received during submission.

Acquisition workflow

Negotiation

Negotiating the terms of the deposit should take place before any records have been transferred. Many aspects of the deposit agreement may be covered in the acquisition policy of the organisation but details about each deposit, especially for local and specialist archives may be required at a collection level. The depositor should state if there are any limitations in what or when the records can be published, for example can some material be opened immediately while others can only be opened on the death of the depositor or after a set period of time?

A key consideration is the right of the organisation to alter the record for preservation purposes, for example migration to a format that can be preserved in the long term or accessed.

If the transfer includes content that is essential to the understanding of the records but does not constitute a record itself there should be an agreement that the organisation can delete those files when their content has been captured for use elsewhere (for example as metadata for the records).

Transfer

Most institutions will need to develop procedures and documents to support the smooth transfer of digital resources from suppliers into their collections.

When transferring born digital objects into an organisation's IT environment consideration should be given as to how this is to take place so as to ensure the security and completeness of the transfer. For smaller organisations it may be sufficient to have the relevant digital files delivered on a drive or similar hardware and check their contents against the descriptive file manifest. Alternatively organisations may wish to transfer digital files using an in-house FTP, or a paid for third party solution (such as a cloud based file sharing service), to ensure chain of custody.

The table below outlines options for transfer and accessioning of digital materials. Decisions on file formats and if relevant storage media (see Storage, Legacy media, and File formats and standards) will support and be interdependent with this process.

Options for Transfer and Accessioning of File Formats and Storage Media		
Options	**Issue**	**Requirements**
All options		• Policy on File formats and standards. • Preservation planning and technology watch on developments in Storage and for Legacy media formats.
Limit range of file formats received **Limit range of media received (most cost-effective long-term option)**	• Simplifies management and reduces overall costs. • Depositor may lack resource or expertise to comply. • Wide variety of file formats used and proprietary extensions to open standards. • Physical storage media used for transfer may only be temporary carriers and content will be transferred to long-term storage used.	• Guidelines on preferred file formats. • Degree of influence over the deposit. • Advocacy and Collaboration strategies to achieve desired outcomes. • Guidelines on preferred transfer media and transfer procedures.
Accept file formats as received but convert to standard file format **Accept storage media as received but transfer contents to standard storage used**	• Simplifies management and reduces longer term costs. • May not be technically feasible to convert to standard file format. • It will be necessary to check that accidental loss of data has not occurred.	• Legal compliance, Copyright permissions or statutory preservation rights. • Resources and technical expertise at host institution. • Election of preferred formats.

		• Documentation of native formats to allow conversion. • Integrity checks for conversion process.
Accept and store as received (least cost-effective option long-term, despite lower initial costs)	• Complicates management and increases costs of managing resources over time. • High risk option, particularly if large numbers of digital materials are being collected. • A choice of file formats may be available. That deposited may not be the most suitable for preservation. • Storage media may be of unknown quality and suitability for long-term preservation. • Formats may be obsolete or not supported within the institution.	• Clearly defined priorities for both short and long-term preservation. • Ability to address issues such as encryption, proprietary software etc. in received items. • Ability to ensure future access to information contained in the item.

Validation

Once transfer has taken place, the files should be located in a secure, quarantined, backed up environment and a check in process should be promptly initiated. Having transferred and housed a copy of a born digital collection an organisation is now liable for certain legal responsibilities such as Freedom of Information Requests if in the public sector. Following this, a letter of acknowledgement of receipt should be sent to the donor. It is important at this early stage that no instruction to destroy the original files is given.

Records should be virus checked at the earliest opportunity to ensure that the material has not been infected with malware or viruses. If any are found the depositing organisation should be alerted and the media either returned to them (if they do not have a copy) or formatted and returned or destroyed according to the depositor's preference. Once the records are confirmed as virus free a check should be carried out to ensure that all the records are present and undamaged. The most reliable method for this is to verify the files against the manifest. Create checksums for the files and compare them against those listed on the manifest pre-transfer. If the checksums match you can be sure that the records have not been corrupted or accidentally altered between the points of transfer from the depositor and arrival at the organisation. If no verifiable manifest was provided with the deposit, it may be impossible to comprehensively verify the integrity of the files and manual viewing of a sample of files may be necessary to provide some indication of completeness and quality. In this case, a verifiable manifest should be generated to enable subsequent fixity checking.

At this stage the records will hopefully have been confirmed as complete (according to the manifest), retaining their integrity (exactly what the depositor supplied) and are virus and malware free. They can now be ingested into the digital preservation system.

Metadata describing the deposited material will assist in ensuring the fixity (see Fixity and checksums) of the material during the transfer process as well as supporting subsequent preservation and access. This might include:

A verifiable manifest consisting of a list of the file and folder names and checksums/fixity values for each file

The size of the files (with a total volume)

A list of the file formats

A statement detailing any IPR associated with the records

Where possible the onus of providing information on the IPR of the records should reside with the depositor.

Ingest Process

The period between transfer to the organisation and ingest into the organisation repository or digital preservation environment may be substantial. This accession phase can be especially prolonged for large born digital collections, sometimes amounting to years, but it is during this phase that a qualitative appraisal of the objects can be made. Items are examined, their technical metadata harvested, their descriptive metadata enhanced and the general accession processes of the organisation applicable to any object take over.

It is during this sometimes prolonged appraisal period that items can be reconsidered for ingest, or rejected if on examination it is felt they do not meet the acquisition or collection profile of the organisation, the file format specification laid out in the guidance documents, or for any other reason. A moratorium may be imposed on items of particular sensitivity such as personal information, commercially sensitive information, or items that break libel laws for instance. In such cases it is important to clearly specify the closure period of the file.

Ingest Procedures to prepare data and documentation for storage and preservation
Unique numbering Each digital resource accessioned by an institution should be allocated a unique identifier. This number will identify the resource in the Institution's catalogue and be used to locate or identify physical media and documentation. In the event of a resource being de-accessioned for any reason, this unique number should not be re-allocated. See Persistent identifiers for advice if you use a persistent identifier scheme.
Handling and Storage transfer guidelines Handling and transfer guidelines for accessioning staff should be developed reflecting IT and preservation staff advice on best practice for different storage media and file transfer to long-term storage systems (see Legacy media, Digital forensics, and Storage).
Re-formatting file formats Where the file formats used to transfer the resource are unsuitable for long-term preservation, the Institution may re-format the resource onto its preferred file formats. In addition to archive

formats, versions in other formats suitable for delivery to users may also be produced from the original (see File formats and standards, and Storage).

Copying

Multiple backup copies of an item may be generated during accessioning as part of institutions' storage and preservation policy and to enable disaster recovery procedures (see Storage).

Security

System and physical security policies and procedures should be in place to ensure the care and integrity of items during accessioning. These should be developed from and reflect the institutional policies and procedures on security (see Information security).

Edition and version control

Procedures for updating and edition control of any dynamic digital materials accessioned (e.g. annual snapshots of databases which are regularly being updated) or for version control of accessioned items where appropriate (e.g. items accessioned in different formats or for which different formats for preservation and access have been generated.)

Cataloguing and documentation standards

Metadata and documentation received or created during transfer, validation and ingest is essential in order effectively to exchange information and documents between platforms and individuals. At a minimum, it should provide information about an item's provenance and administrative history (including any data processing involved since its creation), content, structure, and about the terms and conditions attached to its subsequent management and use including IPR rights and the period over which they pertain (see Metadata and documentation). It should be sufficiently detailed to support:

- Resource discovery (e.g. the location of a resource which is at least briefly described along with many other resources).
- Resource evaluation (e.g. the process by which a user determines whether s/he requires access to that resource).
- Resource ordering (e.g. that information which instructs a user about the terms and conditions attached to a resource and the processes or other means by which access to that resource may be acquired).
- Resource use (e.g. that information which may be required by a user in order to access the resource's information content).
- Resource management (e.g. administrative information essential to a resource's management and preservation as part of a broader collection and including information about location, version control, etc).

Processing times

Ideally targets should be set and monitored for the maximum time between acquisition and cataloguing to prevent backlogs of unprocessed and potentially at risk materials developing during the accessioning process.

Skills, resources and capacity

Organisations should consider whether they have sufficient technical and staffing resources to acquire digital collections. This information may however not be apparent at the outset of an acquisition, as the various challenges of curation of specific digital collections may only reveal themselves over time. Organisations should therefore plan for knowledge, skill and staffing gaps and where possible address these through training, recruitment or engagement of specialist professional digital curation services. Where funding resources cannot meet these, often a dedicated in-house knowledge building drive may suffice for the interim. (see Staff training and development, and Procurement and third party services).

Costs of acquisition and ingest

Trying to establish indicative costs for digital preservation activities is always problematic. These should not just include storage (the most obvious) but should also look at the cost of the staff time required to manage the accession and ingest of each born digital object, a process which can mirror time-wise the accession pathway of physical artefacts. Other anticipated costs might include curation processes like normalisation, analysis, enrichment of metadata, increased robustness of storage, disaster recovery etc.

Although an organisation should find best value solutions for these lifecycle costs, it should be recognised that the investment needed to provide a robust preservation pathway that can safeguard our digital heritage may be significant, and that certain processes must be adhered to irrespective of the nature of an accessioned object. Organisations should therefore bear this in mind when acquiring born digital collections. (see Business cases, benefits, costs, and impact)

Summary of recommendations

Acquisition and appraisal - recommendations checklist

Agreements and Guidance for depositors

- [ ] Create a suite of standard depositor agreements and licences
- [ ] Create appropriate guidance for depositors

Transfer procedures

- [ ] Provide documentation to guide and support transfer of digital materials from suppliers
- [ ] Decide how your transfer procedures can best be developed to support your storage and preservation policies

Validation procedures

- [ ] Check media, content, and structure

Procedures to prepare data and documentation for storage and preservation

- [ ] Unique numbering of each item accessioned
- [ ] Handling and storage transfer guidelines for different media
- [ ] Re-formatting of file formats if required according to agreed guidelines
- [ ] Generating multiple copies of an item as part of an institution's storage and preservation policy
- [ ] System and physical security policy and procedures for items during accessioning

Procedures for cataloguing and documentation

- [ ] A minimum standard of information required for cataloguing including IPR information
- [ ] Guidelines for retrospective documentation or catalogue enhancement.
- [ ] Procedures for updating, and managing versions or editions of an item.
- [ ] Procedures to update collection management databases
- [ ] Selection of cataloguing and documentation standards
- [ ] Targets for accessioning tasks and timescales for their completion

Review of procedures

- [ ] Guidelines and schedules should ideally be reviewed annually, or as often as is practical to keep pace with an organisation's developing requirements and collections development policies

Staff training

- [ ] Plan for knowledge, skill and staffing gaps and where possible address these through training, recruitment or engagement of specialist third-party services

Costs

- [ ] Evaluate and plan for lifecycle costs of acquisitions

Resources

ISO 20104:2015 Space data and information transfer systems -- Producer-Archive Interface Specification (PAIS)

CCSDS 651.1-B-1, Producer-Archive Interface Specification (PAIS) (2014) RECOMMENDED STANDARD CCSDS 651.1-B-1 BLUE BOOK February 2014

http://public.ccsds.org/publications/archive/651x1b1.pdf

The Blue Book is a free to access pre-print of ISO 20104:2015. The PAIS standard aims to provide a standard method for formally defining the digital information objects to be transferred by an information Producer to an Archive and for effectively packaging these objects in the form of Submission Information Packages (SIPs). This supports effective transfer and validation of SIP data (104 pages).

What is appraisal?

http://www.nationalarchives.gov.uk/documents/information-management/what-is-appraisal.pdf

This guidance from The National Archives applies to UK public records in any format, including paper, digital, audio, film or model format as defined by the Public Records Act 1958, and all organisations responsible for such records (2013, 7 pages).

Preserving eBooks, DPC Technology Watch Report 14-01 July 2014

http://dx.doi.org/10.7207/twr14-01

This report discusses current developments and issues with which public, national, and higher-education libraries, publishers, aggregators, and preservation institutions must contend to ensure long-term access to eBook content and which affect acquisition as well as preservation (31 pages).

Preservation, Trust and Continuing Access for e-Journals, DPC Technology Watch Report 13-04 September 2013

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

This report 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 is not solely focused on technology, and covers relevant legal, economic and service issues in acquiring access to networked digital resources and the unique preservation challenges this presents (43 pages).

The UNESCO/PERSIST Guidelines for the selection of digital heritage for longterm preservation

https://www.unesco.nl/sites/default/files/dossier/persistcontentguidelinesfinal1march2016.pdf?download=1

The UNESCO/PERSIST (Platform to Enhance the Sustainability of the Information Society Transglobally) Project released these Guidelines on the selection of digital heritage for long-term preservation in March 2016. The aim of the Guidelines is to provide an overarching starting point for libraries, archives, museums and other heritage institutions when drafting their own policies on the selection of digital heritage for long-term sustainable digital preservation. (19 pages).

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 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. You can use the POWRR Tool Grid to see which technical tools in COPTR can help to support acquisition, ingest, or multiple functions.

DLF/OCLC Registry of Digital Masters

http://www.diglib.org/community/groups/rdm/

The DLF/OCLC Registry of Digital Masters provides a central place for library staff to search for, and find, digitally preserved materials. Typical items include digitised monographs and serials. A registered object ensures that the digital object (or soon to be digitised) followed established standards and best practices for digitisation and that the institution that digitised it has made a commitment to digital preservation of this object.

Keepers Registry

http://thekeepers.org

The Keepers Registry acts as a global monitor on the archiving arrangements for electronic journals. It has three main purposes to: enable librarians and policy makers to find out who is looking after which e-journal, how and with what terms of access.; highlight the e-journals which are still "at risk of loss"; and showcase the organisations (the keepers) which act as digital shelves for access over the long term. It has a Title List Comparison feature to help you discover the archival status of a list of serial titles important to you: reporting those which are being archived and those which are "at risk".

MediaRIVERS (Media Research and Instructional Value Evaluation and Ranking System)

https://github.com/IUMDPI/MediaSCORE

Software created by Indiana University in collaboration with AVPreserve guides a structured assessment of research and instructional value for media holdings. The free, open source version requires installation and configuration on a server, and a hosted application is available on a monthly subscription basis.

Practical E-Records:software and tools for archivists

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

Pages created by Chris Prom for Transfer Guidelines, E-Records Deposit Policy, and Submission Agreement Form provide sample templates that you can modify and/or provide to record producers whose records your repository wishes to accession. Permission to modify and republish these transfer guidelines is provided under a Creative Commons Attribution 3.0 United States License.

Archaeology Data Service Guidelines for Depositors

http://archaeologydataservice.ac.uk/advice/guidelinesForDepositors

The ADS Guidelines for Depositors provide guidance on how to correctly prepare data and compile metadata for deposition with ADS and describe the ways in which data can be deposited. There is also a series of shorter summary worksheets and checklists covering: data management; selection and retention; preferred file formats and metadata. Other resources for the use of potential depositors include a series of Guides to Good Practice, which complement the ADS Guidelines and provide more detailed information on specific data types.

Selecting and transferring records

http://www.nationalarchives.gov.uk/information-management/manage-information/selection-and-transfer/

These pages provide guidance on the selection and transfer of records. UK bodies transferring records to The National Archives or to places of deposit under the Public Records Act 1958 should follow this process for records in all formats and media, including paper and digital records. It consists of guidance on six steps:

Step 1: Appraising your records

Step 2: Selecting your records

Step 3: Sensitivity reviews of selected records

Step 4: Cataloguing and preparation of records

Step 5: Planning and arranging delivery of records

Step 6: Accessioning your records

Case Studies

The Work of Appraisal in the Age of Digital Reproduction

http://archival-integration.blogspot.co.uk/2015/06/the-work-of-appraisal-in-age-of-digital.html#pii

The Bentley Historical Library's ArchivesSpace-Archivematica-DSpace Workflow Integration project discussion highlights current digital archives appraisal techniques employed by the Bentley, many of which they are hoping to integrate into Archivematica (June 2015).

Acquisition & management of digital collections at the Library of Congress

http://www.slideshare.net/NASIG/acquisition-management-of-digital-collections-at-the-library-of-congress-34244613

The Library of Congress, as the national library and the home of the US Copyright Office, is heavily involved in digital acquisition and management. This concise and informative powerpoint by Ted Westervelt shares the experiences that the Library of Congress has had and lessons it has learned (2014, 30 slides).

Trust Me, I'm an Archivist: Experiences with Digital Donors

http://www.ariadne.ac.uk/issue65/hilton-et-al

This 2010 article by staff at the Wellcome Trust Library discusses four common scenarios that seem to act as new blocks to the transfer of digital material: Lack of Long-term Planning; IT vs Records Management; Duplication and Abundance; and The Fear of Digital. It concludes that we need to change the way we present information, how we work with digital material and how we can support and assist our donors. The degree of engagement that is standard practice with paper records will not suffice for born-digital material: our interaction with depositors will ideally be even closer and even more frequent, as we help them deal not merely with new technical challenges but with the plethora of soft-skills issues, of preconceptions and of attachments that surround them.

Retention and Review

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Selection for long-term retention will normally occur at acquisition but can be an iterative process occurring at later stages once an item is already in the collections. The term retention and review is used here for this iterative process. The decision process mirrors steps included in the Decision Tree in Acquisition and appraisal and the tree can be adapted for this purpose.

Employing evaluation criteria and selection procedures for all potential digital acquisitions ensures that collections development is carefully prioritised and sustainable. The use of such criteria and procedures should minimise the frequency and need for retention and review decisions, as acquisitions are carefully evaluated and justified prior to entering the collections. Organisations may also need to retain certain in house records and digital materials for regulatory, legal, operational and financial requirements. These should similarly be actively managed to retain their viability, authenticity, and accessibility.

Digital items acquired over time and before institutional policies and procedures were in place will normally require review. This may be one of the first steps that an institution undertakes in implementing a digital preservation policy: quantifying its current digital holdings and assessing preservation risks (see Getting started).

Archives use the series concept for a body of records that share similar characteristics. Typically, many series are on-going for decades. However, the scope and coverage of a digital series may change over time and certainly technology considerations are likely to change and some attention must be given to a careful evaluation as each accession is transferred to the archives.

Over time the need may also arise to review collections and collections policy to reflect changing needs and circumstances. The necessity of making early decisions on selection for preservation in a digital environment (without the period of hindsight which is often available in analogue environment) may mean that future review may be necessary in the preservation life cycle of digital materials.

In a digital library environment where collection levels have been employed, digital materials in any collection level category can be subject to periodic review, re-designated from one level to another, withdrawn, or de-accessioned as required to meet changing needs and circumstances. However, for items selected for permanent preservation it is anticipated that review and de-accessioning will occur in rare and strictly controlled circumstances. For other collection levels such as mirrored or licensed resources review criteria may include:

- A sustained fall of usage to below acceptable levels.
- The availability of content elsewhere to a higher degree of quality or at considerably lower cost.

Content that has been superseded or is no longer sufficiently accurate to justify maintenance in active form. In such cases, the content may be retained together with subsequent editions or withdrawn.

- Expiry or termination of a licence or data exchange agreement and withdrawal/return of a digital resource to the data supplier.
- Cost to sustain the data resource outweighs the value/benefit received.
- Deterioration in the quality service provided by a supplier or deterioration in the accessibility of content due to poor updating of indexing, imaging, or other characteristics internal to the data resource.

Within the archives and records management professions the use of retention periods and schedules is well established. Records may be destroyed at the end of their retention period, retained for a further period, or transferred to an institution for long-term preservation.

In any collection environment it is important that written procedures are in place for the process of retention and review with clear responsibilities assigned to named individuals or those sections of an organisation in charge of governance and collections development. The timescales, circumstances, and authorisation procedures for the review should be clearly stated. Retention and review schedule documents themselves should be periodically reviewed to keep pace with emerging organisational requirements. Depending on the institution's business environment, its users and depositors may be consulted as part of the process. Any recommendations may then be referred for approval to management and committees as appropriate to the size and significance of the resource.

Where a recommendation is made to de-accession an archived resource there should be procedures to consult with other stakeholders to determine whether transfer to another organisation should occur. In such cases the institution should agree conditions of transfer which include acceptable levels of care for the resource and access to it as appropriate for educational and research users.

Financial constraints should not be a main driver for de-accessioning digital objects as the process itself is not without cost and may raise thorny ethical issues. Decisions to de-accession should primarily be driven by the collections development policy.

Accessioned digital materials that have not been retained after review, should retain their entry in any institutional catalogue with comments identifying the process undertaken and any transfer details.

Succession planning by an institution may also be relevant to retention and review. The standard for Audit and Certification of Trustworthy Digital Repositories (CCSDS, 2011) recommended practice 3.1.2.1 requires that the repository shall have an appropriate succession plan, contingency plans, and/or escrow arrangements in place in case the repository ceases to operate or the governing or funding institution substantially changes its scope.

Resources

Deaccessioning and disposal: Guidance for archive services

http://www.nationalarchives.gov.uk/documents/Deaccessioning-and-disposal-guide.pdf

This guide was written to support Archive Service Accreditation, the UK-wide standard for archive services. It is generic and applies to content in all formats, analogue or digital and has no special provisions for digital materials. The standard presents de-accessioning as part of collections development, and requires archive services to have policies, plans and procedures in place for collections development activities including de-accessioning. It includes a disposal destination decision tree (2015, 34 pages).

Digital Preservation Management tools: Digital Content Review process

http://dpworkshop.org/workshops/management-tools/process-results

To complete a digital content review, the digital preservation team gathers information and iteratively accumulates as part of a structured process. The results of ongoing digital content reviews produce a digital content review dataset that enables near-term and long-term planning by organizations.

The National Archives, Disposing of records

http://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/disposal/

TNA have produced step-by-step guidance to help you through the disposal process including a disposal checklist.

The British Library's de-accessioning policy

http://www.bl.uk/aboutus/stratpolprog/coldevpol/deaccessioning/

This policy sets out the circumstances under which the British Library may dispose of certain types of material. It is a generic policy that applies to content in all formats analogue or digital and has no special provisions for digital materials.

References

Consultative Committee for Space Data Systems (CCSDS), 2011. *Audit and Certification of Trustworthy Digital Repositories Recommended Practices CCSDS 652.0-M-1 Magenta Book September 2011* [this was subsequently published as ISO 16363: 2012]. Available: http://public.ccsds.org/publications/archive/652x0m1.pdf

Storage

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Introduction

This section covers the emerging practice of using IT storage systems for digital preservation. It deals with generic issues and more specific issues associated with cloud storage are addressed separately (see Cloud services section). The traditional practice of preserving of digital media, for example legacy items within existing collections is also covered elsewhere (see Legacy media section). Many organisations will have mixed strategies or will be in the process of transitioning from one to the other.

The use of storage technology for digital preservation has changed dramatically over the last twenty years. During this time, there has been a change in practice. Previously, the norm was for storing digital materials using discrete media items, e.g. individual CDs, tapes, etc., which are then migrated periodically to address degradation and obsolescence. Today, it has become more common practice

to use resilient IT storage systems for the increasingly large volumes of digital material that needs to be preserved, and perhaps more importantly, that needs to be easily and quickly retrievable in a culture of online access. In this way, digital material has become decoupled from the underlying mechanism of its storage. With this come consequent benefits of allowing different preservation activities to be handled independently.

Resilient storage systems

A resilient IT storage system consists of storage media contained within a server that provides built in resilience to various failure modes by using inbuilt redundancy and recovery. For example, a storage system might be hard disk drives in a Redundant Array of Independent Disks (RAID), data tapes in a tape library, or a combination of storage types in a Hierarchical Storage Management system (HSM). It can include onsite storage and/or remote cloud storage and automated replication of digital materials across multiple sites and systems.

These systems will still become obsolete over time and digital materials should be migrated regularly between storage systems as they become obsolete. Migration between storage systems is separate to migration between file formats and can be handled largely as an IT issue, with the proviso that proper oversight is employed to ensure preservation requirements are met. The upside is that the use of IT systems for data storage can provide much faster access, a more scalable solution, easier management, and ultimately lower costs especially at scale.

It is critical to understand the difference between standard IT storage solutions and the additional needs of long-term preservation. It is essential to be able to explain these differences to your IT department or storage service provider and to be able to specify these requirements when procuring a system or service. Standard storage systems are designed for digital objects that are in active use. While backup procedures are usually included, they generally do not meet the more stringent requirements to ensure long-term preservation of digital materials. Backup and digital preservation are not the same thing and many IT departments or experts may not appreciate this. Preservation storage systems require a higher level of geographic redundancy, stronger disaster recovery, longer-term planning, and most importantly active monitoring of data integrity in order to detect unwanted changes such as file corruption or loss.

There are many ways of meeting requirements for preservation storage and these will vary in scale and complexity depending on organisational context. It will be necessary to assess in-house resources and consider out-sourcing and cloud storage options. The approach taken will often depend on the collection size and complexity of collection and the resources that are available within the organisation. It is possible to meet preservation storage requirements with a basic set-up, but as a collection increases in size it will be necessary to address issues such as scalability and automation.

	Principles for using IT storage systems for digital preservation
	The following represent principles that should be employed when designing or selecting storage systems for preservation.
1	**Redundancy and diversity** • Make multiple independent copies of digital material and store these in different geographic locations.

	• Use a combination of online storage systems and offline media. • Use different types of storage technology to spread risk and achieve a balance of data safety, easy access and manageable cost.
2	**Fixity, monitoring, repair** • Use fixity measures such as checksums to record and regularly monitor the integrity of each copy of the digital material. • If corruption or loss is detected then use one of the other copies to create a replacement. • Store fixity information alongside the digital materials and also in separate databases or systems.
3	**Technology and vendor watch, risk assessment, and proactive migrations** • Understand that storage technologies, products and services all have a short lifetime. • Use technology watch to assess when migrations might be needed. • Keep an eye on the viability of storage vendors or classes of storage solution. • Be proactive in migrating storage before digital material becomes at risk
4	**Consolidation, simplicity, documentation, provenance and audit trails** • Minimise the proliferation of legacy media types and storage systems used for preservation. • Consolidate digital materials onto the minimum number of preservation storage systems (subject to the redundancy requirements above). • Document how digital materials have been acquired and transferred into the storage systems as well as how the storage systems are set-up and operated. • Use this to provide audit information on data authenticity.

Storage reliability

When looking at storage solutions, either onsite or cloud, the question arises of how reliable they are and if something goes wrong then what does this mean in terms of data loss. Manufacturers will typically assert statistics such as reliability, durability, failure rates and error rates.

For example, this might take the form of a cloud service being designed for 99.999% durability, a Bit Error Rate (BER) of 1 in 10^{16} when reading data from a data tape, or a Mean Time Between Failure (MTBF) of 1M hours for a hard drive. These numbers are then used to calculate further measures of reliability. For example, a Mean Time To Data Loss (MTTDL) of 1,000 years might be asserted when hard drives are used in a RAID6 array.

These numbers can be hard to understand, and they need to be interpreted with great care when attempting to estimate 'how safe' a given storage solution will be.

There has already been substantial work on how to describe, measure and predict storage reliability, including from a digital preservation perspective. This is a complex topic that is not possible to cover

in this handbook. Some example references for further reading are Greenan et al (2010), Rosenthal (2010) and Elerath (2009). What comes from this work are several important considerations:

- IT Storage technology is in general remarkably reliable for what it does. Failures are relatively rare events, but they do and will happen. The temptation is to assume that just because at an individual level a particular type of failure hasn't been experienced then that storage technology is in general more reliable than it really is. This is dangerous position. For example, many people will have hard drives that have worked perfectly well for years and years, but the reality is that, on average, up to 5-15% of hard drives actually fail within one year (Backblaze, 2014), (Pinheiro et al, 2007).

- Because failures and errors are relatively rare events, reliability statistics from vendors are typically based on models and simulations and not from long-term observations of what actually happens in practice. For example, if a manufacturer says that the shelf life of media is 30 years then it's not because they have actually tested media over that time period. Likewise, if a vendor estimates the MTTDL is 1,000 years then they clearly haven't built a system and tested it for anywhere near that length of time. Therefore, statistics should be interpreted as best estimates from vendors of how a system might behave in practice - but it may not actually turn out that way. For example, field studies have suggested that manufacturer estimates of reliability can be over optimistic (Jiang et al, 2008) .

- The likelihood of data loss increases dramatically when correlations are taken into account. Correlations are where parts of a system, or different copies of the digital material, can't be considered as independent. If there is a problem with one part of the system or copy of the digital material then there is likely to be a problem with another part or copy. Examples include a manufacturing fault affecting all the hard drives in a storage server, software or firmware bugs systemically corrupting digital material, failure by an organisation to regularly test its backups, or failure to isolate or decouple storage systems so that if one copy of the digital material is accidentally deleted then all the other copies don't get deleted too. These correlation factors can be far more significant than the specific failure modes covered by reliability statistics.

These findings and observations result in the following recommendations:

- Plan for failures to happen in IT storage solutions no matter how cleverly designed by the manufacturer. Failure rates in practice may well be higher than manufacturer statistics suggest.

- Data loss can be caused by failure to put in place proper processes and procedures around the use of IT storage as well as from the storage technology itself. Proper risk assessment is the way to identify and manage these problems.

- The best strategy remains to create multiple independent copies of digital material in different locations and to store them using different technologies where possible. This should include a process of actively and regularly checking data integrity of all the copies so problems can be detected no matter why and where they might occur. In this way, risks are both minimised and spread, and reliance isn't placed on any particular storage technology or service being completely error free.

Multi-copy storage strategies

Digital storage technologies present several risks to long-term preservation of digital objects. These risks can be reduced by using a digital storage strategy that involves one or more storage systems and at least two copies of the data.

Good practice is for a storage strategy to have the following characteristics:

(a) multiple independent copies exist of the digital materials

(b) these copies are geographically separated into different locations

(c) the copies use different storage technologies

(d) the copies use a combination of online and offline storage techniques

(e) storage is actively monitored to ensure any problems are detected and corrected quickly.

A digital storage strategy can be implemented in a staged way, starting with a basic level of protection and access to digital content and moving on towards a more automated and scalable approach that gives a higher level of data safety and security.

Risks to digital content come from a range of sources and a digital storage strategy helps balance the cost of digital storage with the reduction of those risks. Example risks to consider include fire, flood, failure to instigate or follow proper processes or procedure, malicious attack, media degradation, and obsolescence of storage systems and technologies. The principal risks and means of addressing or mitigating them are often addressed in an organisation's business continuity planning (see Risk and change management).

It is important to realise that many examples of content loss are not necessarily due to technical faults with storage technology (although it is important to recognise that these do happen), but can come from human error, lack of budget or planning of storage migrations, or a failure to regularly check and correct failures that might occur.

In a world that is increasingly using networked systems and technologies for digital storage, there is a role for an offline copy of digital materials. This can provide a 'fire break' against problems with online systems that can automatically propagate between locations, e.g. deletion of a file in one location that automatically deletes a mirrored copy at another site.

Making more than one copy of the digital materials is fundamental to achieving a basic level of data safety. Using different types of storage for each copy helps spread the risk and ensure that a problem with one technology doesn't affect the others. The way each copy is stored can be adjusted to achieve an acceptable overall level of cost, risk and complexity. For example, one copy might be held using an online storage server for fast access and one copy might be on data tape in deep archive for low cost and relatively high safety.

This Handbook follows the National Digital Stewardship Alliance (NDSA) preservation levels (NDSA, 2013) below in recommending four levels at which digital preservation can be supported through storage and geographic redundancy. We make the additional recommendation of using a combination of online and offline copies to achieve a good combination of data access and data safety:

Level	Approach	Risks addressed and benefits achieved
1	• Two complete copies of the digital materials that are not co-located. One copy should be offline. • For digital materials on heterogeneous media (optical disks, hard drives, etc.) get the content off the medium and into your storage system.	• Basic ability to recover from a range of issues including storage system failure. Loss or damage to one copy can be recovered using the other copy. • Digital materials easier to manage when in a single storage system.
2	• At least three complete copies. • At least one copy in a different geographic location. • Document your storage system(s) and storage media and what you need to use them.	• As above plus protection from natural disasters and other major events. • Good level of access and digital materials safety. • Staff have clear policies and procedures to follow so are more efficient, costs are lowered, and staff changes can be managed.
3	• At least one copy in a geographic location with a different disaster threat. • Obsoloscence monitoring and migration process for your storage system(s) and media.	• As above plus protection from the longer-term risks associated with technical obsolescence. • Continual access to content is possible even during migrations and disasters.
4	• At least three copies in geographic locations with different disaster threats. • Have a comprehensive plan in place that will keep files and metadata on currently accessible media or system.	• As above with full range of risks addressed including accidental loss and malicious attack, vendor lock-in, and budget instabilities. • Content has high availability, costs are predictable and manageable, there is the ability to achieve trusted repository certification.

Managing storage system obsolescence and risks

The use of storage technologies and solutions needs careful planning and management to be an effective approach to supporting digital preservation. If done properly, the result can be very good levels of data safety, rapid access to content when needed, and costs that are both low and predictable.

IT storage technologies can fail or cause data corruption and the lifetime of media and systems is typically short, for example 3-5 years, which means solutions become obsolete quickly and migration is needed to avoid digital materials becoming at risk. Migration in this context means moving data

off an old storage system and onto a new storage system. The digital material itself does not change but the storage solution does. An IT department or storage service provider will think of migration at the storage level. This is in contrast to file format migration where the file format will change, but the way that the files are stored doesn't change.

Resources

NDSA Levels of Preservation

http://www.digitalpreservation.gov/ndsa/activities/levels.html

The National Digital Stewardship Alliance (NDSA) "Levels of Digital Preservation" are a tiered set of recommendations for how organizations should begin to build or enhance their digital preservation activities. It is intended to be a relatively easy-to-use set of guidelines useful not only for those just beginning to think about preserving their digital assets, but also for institutions planning the next steps in enhancing their existing digital preservation systems and workflows. It is not designed to assess the robustness of digital preservation programs as a whole since it does not cover such things as policies, staffing, or organizational support.

These are some of the more notable digital preservation storage systems and storage system/service providers There are a wide-range of commodity IT storage vendors, as well as specialist digital preservation service providers that can provide onsite or cloud storage (see also Cloud services). These specialists typically may support other preservation functions in addition to storage.

Arkivum

http://arkivum.com

Digital Preservation Network

http://www.dpn.org

DSpace

http://www.dspace.org

ePrints

http://www.eprints.org

Fedora

http://fedorarepository.org

iRods

http://irods.org

LOCKSS

http://www.lockss.org

OCLC Digital Archive

http://www.oclc.org/digital-archive.en.html

Portico

http://www.portico.org/digital-preservation/

Preservica

http://preservica.com

Rosetta

http://www.exlibrisgroup.com/category/RosettaOverview

Community Owned digital Preservation Tool Registry COPTR

http://coptr.digipres.org/Main_Page

Although focussing principally on tools the COPTER registry also covers a range of storage systems and services. It acts primarily as a finding and evaluation tool to help practitioners find the tools they need to preserve digital data. 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.

DSHR's Blog

http://blog.dshr.org

David Rosenthal is a computer scientist and chief scientist for the LOCKSS project. His blog frequently covers computer storage development and trends and implications for digital preservation.

Case studies

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

http://www.nationalarchives.gov.uk/documents/archives/case-study-oxford.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 (4 pages).

The National Archives case study: Parliamentary Archives

http://www.nationalarchives.gov.uk/documents/archives/case-study-parliament.pdf

This case study covers the Parliamentary Archives. It 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 (4 pages).

The National Archives case study: Tate Gallery

http://www.nationalarchives.gov.uk/documents/archives/case-study-tate-gallery.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 (4 pages).

Use Cases: Storage Solutions for Digital Preservation

http://www.alliancepermanentaccess.org/index.php/aparsen/webinars/#09

A set of powerpoint presentations covering use cases and survey results from an APARSEN event on 14 April 2014. The uses cases for the National Library of the Netherlands (KB) and RSS Data Farm (European Space Agency) are particularly relevant.

References

Backblaze, 2014. Hard Drive Reliability Update – Sep 2014. *Backblaze*. [blog] Available: https://www.backblaze.com/blog/hard-drive-reliability-update-september-2014/

Elerath, J., 2009. Hard-Disk Drives: The Good, the Bad, and the Ugly. *Communications of the ACM*. 52 (6), 38-45. Available: doi:10.1145/1516046.1516059. http://cacm.acm.org/magazines/2009/6/28493-hard-disk-drives-the-good-the-bad-and-the-ugly/fulltext

Greenan, K.M., Plank, J.S. & Wylie, J.J., 2010.Mean time to meaningless: MTTDL, Markov models, and storage system reliability. *Proceedings of the 2nd USENIX conference on Hot topics in storage and file systems*. Available: https://www.usenix.org/legacy/event/hotstorage10/tech/full_papers/Greenan.pdf

Jiang, W. et al., 2008. Are Disks the Dominant Contributor for Storage Failures? A Comprehensive Study of Storage Subsystem Failure Characteristics. *Proceedings of the 6th USENIX Conference on File and Storage Technologies (FAST '08)*. Available: http://www.usenix.org/events/fast08/tech/jiang.html

NDSA, 2013. T*he 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

Pinheiro, P., Weber, W-D. & Barroso, L.A., 2007. Failure Trends in a Large Disk Drive Population. *Proceedings of the 5th USENIX Conference on File and Storage Technologies (FAST' 07)*. Available: http://static.googleusercontent.com/media/research.google.com/en//archive/disk_failures.pdf

Rosenthal, D.S.H., 2010. Bit Preservation: A Solved Problem? *The International Journal of Digital Curation*. 5 (1) Stanford University Libraries, CA. Available: http://www.ijdc.net/index.php/ijdc/article/view/151/224

Legacy Media

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Introduction

Many organisations will have large amounts of data stored on legacy media, such as magnetic and optical media, and data will continue to be received on old carriers. Ultimately, the best long-term strategy for the preservation of the data will be migration to file-based storage and active management thereafter (see Storage section). Often the original media will continue to be preserved alongside this, so it will be necessary to understand their preservation and storage requirements. For organisations with large collections of legacy media, understanding the risks facing each media type will also help with prioritising collections for migration and application of digital forensics tools and methods will also be helpful (see Digital forensics section).

For the preservation of magnetic and optical media, two aspects need to be considered - the media itself and the hardware and software needed to interpret it. In some cases the second aspect will be the most challenging. As the popularity of a media format declines, the manufacture of hardware ceases and becomes more difficult to procure and maintain.

Preserving legacy media

In most cases, the simplest way to mitigate risks with storage media is to transfer all content into a managed storage system. This means that the content can be managed without reference to the original storage medium. This would probably be adequate for the vast majority of digital content requiring preservation. However, there may be a few instances where it is necessary to retain the original media carrier in some way. In some cases, the storage medium could simply be retained as an artifact, with no expectation of long-term access, e.g. where it forms part of a hybrid collection or has some kind of value by association. (e.g. part of the collections of a prominent author). However, where continued access to the content is required, careful thought needs to be given to how it could be accessed in the future.

One thing that we do know from experience is that digital storage media types change frequently over time. For example, the previous version of this handbook contained an overview of magnetic and optical storage media and provided estimates of the lifetimes of selected storage media types that were popular in the mid-1990's (a digital preservation handbook written in previous decades would presumably have included assessments of punched cards and paper tape). Given current trends in storage technology, it is perhaps better now to provide a framework that supports the ongoing evaluation of storage media, which might now include flash memory sticks or external hard-drives. One such framework has been provided by The National Archives (Brown, 2008). This uses a scorecard approach, measuring selected storage media against six criteria:

- longevity (e.g., proven operational lifetimes)
- capacity
- viability (e.g., in terms of retaining evidential integrity)
- obsolescence
- cost
- susceptibility (e.g., to physical damage and to different environmental conditions).

In practice, however, these kinds of assessment can only get you so far. There is a growing body of evidence that suggests that variation in manufacturing quality also plays a major role in media longevity (Harvey, 2011). That is why, in the end, digital preservation normally depends upon the transfer of content from media into a managed storage environment.

Resources

Selecting storage media for long-term preservation, TNA Digital Preservation Guidance Note 2: August 2008

https://www.nationalarchives.gov.uk/documents/selecting-storage-media.pdf

This document is one of a series of guidance notes produced by The National Archives,giving general advice on issues relating to the preservation and management of electronic records. It is intended for use by anyone involved in the creation of electronic. It provides information for the creators and

managers of electronic records about the selection of physical storage media in the context of long-term preservation. Note guidance is as of August 2008. (7 pages).

Care, Handling and Storage of Removable media, TNA Digital Preservation Guidance Note 3: August 2008

http://www.nationalarchives.gov.uk/documents/information-management/removable-media-care.pdf

This document is one of a series of guidance notes produced by The National Archives,giving general advice on issues relating to the preservation and management ofelectronic records. It provides advice on the care, handling and storage of removable storage media. Note guidance is as of August 2008. (10 pages).

You've Got to Walk Before You Can Run: First Steps for Managing Born-Digital Content Received on Physical Media

http://www.oclc.org/content/dam/research/publications/library/2012/2012-06.pdf

A step by step guide about getting digital born material off of various physical media. It focuses on identifying and stabilizing your holdings so that you'll be in a position to take additional steps as resources, expertise, and time permit. The POWRR project document **Resources for Technical Steps** (*3 pages*) adds additional resources for some of the steps. (7 pages).

Kryoflux: Commercial tool for reading floppy disks

http://www.kryoflux.com/

KryoFlux is a USB-based device designed specifically to acquire reliable low-level reads suitable for software preservation. This is the official hardware developed by The Software Preservation Society,

Digital Preservation Management: Chamber of Horrors

http://dpworkshop.org/dpm-eng/oldmedia/disks.html

Some examples of obsolete and endangered disks.

Lost Formats

http://www.experimentaljetset.nl/archive/lostformats

Web page from the Lost Formats Preservation Society with a very nice overview of silhouettes of the shapes to allow quick identification and key brief history and features such as dimensions and storage capacity. All silhouettes shown as same size rather than to scale. Last major update appears to be c.2008 but content is still valuable for all but the most recent formats.

Museum Of Obsolete Media

http://www.obsoletemedia.org/category/format/

Great resource covering a very wide-range of obsolete audio, video, data, and film storage media. You can browse the categories or the Gallery and Timeline. Particularly good if you know what you are looking for and derived mostly from the relevant Wikipedia entries.

Case studies

A Fistful of Floppies: Digital Preservation in Action

https://ischool.uw.edu/sites/default/files/capstone/posters/JStanley_Capstone_Landscape.pdf

The University of Washington Library system currently holds a small collection of electronic thesis and dissertation (ETD) accompanying materials from the late 1980's to 2011 on floppy disks and CD-Rs. These materials will soon reach or have already exceeded the limit of their expected lifespans. This 2015 project looked at the digital preservation possibilities for this collection of materials using digital forensics as a model.(1 page).

Enford, D., et al 2008, Media Matters: developing processes for preserving digital objects on physical carriers at the National Library of Australia, Papers from 74th IFLA General Conference and Council

http://archive.ifla.org/IV/ifla74/papers/084-Webb-en.pdf

The National Library of Australia had a relatively small but important collection of digital materials on physical carriers, including both published materials and unpublished manuscripts in digital form. The Digital Preservation Workflow Project aimed to produce a semi-automated, scalable process for transferring data from physical carriers to preservation digital mass storage, helping to mitigate the major risks associated with the physical carriers. (17 pages).

Digital Preservation Planning Case Study

http://www.dpconline.org/component/docman/doc_download/863-2013-may-getting-started-london-planning-case-study-ed-fay

Presentation on getting started with digital preservation planning, including scoping, risk assessing and prioritising your collection (including legacy media examples), and staff roles and responsibilities. 2013 (20 pages).

References

Brown, A., 2008. Selecting storage media for long-term preservation. *TNA Digital Preservation Guidance Note 2: August 2008*. Available: https://www.nationalarchives.gov.uk/documents/selecting-storage-media.pdf

Harvey, R., 2011. *Preserving Digital Materials 2nd edition*. De Gruyter Saur.

Preservation Planning

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

What is preservation planning?

Preservation planning is the function within a digital repository for monitoring changes that may impact on the sustainability of, or access to, the digital material that the repository holds. It should be proactive: both current and forward-looking in terms of acquisitions and trends. Changes might occur within the repository, within the organisation in which the repository resides, or external to the repository and organisation themselves. Changes might be monitored in the following areas:

- Technology watch
    - Packaging
    - Storage
    - Formats
    - Tools
    - Environment
    - access mechanisms
- Designated communities
    - needs and expectations of users
    - needs and expectations of producers
    - emerging tools for machine to machine access
    - formal feedback from users and producers

The concept of preservation planning is defined within the functional model of the OAIS standard (CCSDS, 2012). This section focuses primarily on the Monitoring components within the OAIS definition. The 'Monitor Technology' and 'Monitor Designated Community' functions of OAIS provide surveys that inform preservation planning activities. These alert the repository about changes in the external environment and risks that could impact on its ability to preserve and maintain access to the information in its custody, such as innovations in storage and access technologies, or shifts in the scope or expectations of the Designated Community (see Lavoie, 2015,13). Preservation planning then develops recommendations for updating the repository's policies and procedures to

accommodate these changes. The Preservation planning function represents the OAIS's safeguard against a constantly evolving user and technology environment. It detects changes or risks that impact the repository's ability to meet its responsibilities, designs strategies for addressing them, and assists in the implementation of these strategies within the archival system.

What is the purpose of preservation planning?

Identifying triggers for taking action to preserve digital materials

Where change has been identified, a risk assessment process can be used to analyse and identify the change that represents a significant risk to the digital material in the repository. Risks can then be addressed and hopefully mitigated following a preservation planning exercise to decide on appropriate preservation action. In this case, the monitoring or technology watch process is identifying trigger points for further analysis, preservation planning and, where relevant, action to preserve digital materials.

Building a knowledge base to inform preservation activities

The process of monitoring internal and external factors as part of a preservation planning activity can inform the knowledge base of an organisation, and in doing so improve its ability to perform digital preservation activities effectively. For example the "knowledge base" of an organisation might be augmented with information about the capabilities of a new software tool, or the obsolescence and unavailability of an existing tool. In some cases this process might be best performed individually or within an organisation, but alternatively might be more usefully performed in a collaborative manner. The vast depth and breadth of knowledge required for digital preservation naturally favours a collaborative approach, whereby particular organisations are able specialise in a particular area and contribute that knowledge to an open or shared knowledge base.

Implementations of a preservation planning service

The degree to which technology watch will be necessary will vary according to the degree of uniformity or control over formats and media that can be exercised by the institution. Those with little control over media and formats received and a high degree of diversity in their holdings will find this function essential. For most other institutions the IS strategy should seek to develop corporate standards so that everybody uses the same software and versions and is migrated to new versions as the products develop.

Failure to implement an effective technology watch or IS strategy incorporating this will risk potential loss of access to digital holdings and higher costs. It may be possible for example to re-establish access through digital forensics (see Digital forensics) but this may be expensive compared to pre-emptive strategies.

A retrospective survey of digital holdings (see Getting started) and a risk assessment and action plan (see Risk and change management) may be a necessary first step for many institutions, prior to implementing a technology watch.

Good preservation metadata in a computerised catalogue identifying the storage medium, the necessary hardware, operating system and software will enable a technology watch strategy (see Metadata and documentation).

Integrated preservation systems, and individual tools and registries can also support this function (see Technical solutions and tools).

Resources

Some of the core preservation watch activities are generic and therefore ready made for collaboration while others are highly localised and not easily shared.

DPC Technology Watch Report Series

http://www.dpconline.org/advice/technology-watch-reports

These reports provide an advanced introduction to specific issues for those charged with establishing or running services for long term preservation and access. They are updated and new reports added periodically.

Scout – a preservation watch system, OPF blog post 16th Dec 2013

http://openpreservation.org/blog/2013/12/16/scout-preservation-watch-system/

The SCAPE Project designed a demonstrator for an automated preservation watch service, called SCOUT. SCOUT was described by its developers as providing "...an ontological knowledge base to centralize all necessary information to detect preservation risks and opportunities. It uses plugins to allow easy integration of new sources of information, as file format registries, tools for characterization, migration and quality assurance, policies, human knowledge and others."

Assessing file format risks: searching for Bigfoot? OPF Blog post 29th Oct 2014

http://openpreservation.org/blog/2013/09/30/assessing-file-format-risks-searching-bigfoot/

This detailed blog post raises concerns about challenges with automating preservation watch.

Barbara Sierman, Paul Wheatley 2010 Evaluation of Preservation Planning within OAIS, based on the Planets Functional Model Planets Deliverable no. PP7-D6.1

http://www.planets-project.eu/docs/reports/Planets_PP7-D6_EvaluationOfPPWithinOAIS.pdf

The Planets Project realised various aspects of the concepts defined within the OAIS Preservation Planning function, and performed an evaluation of OAIS based on these practical experiences. 2010 (34 pages).

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, and Preserving digital Objects With Restricted Resources project (POWRR). 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.

Case studies

OCLC Research Report - Preservation Health Check: Monitoring Threats to Digital Repository Content

http://www.oclc.org/research/themes/research-collections/phc.html

The OCLC Research Preservation Health Check activity was initiated by Open Planets Foundation. The Pilot used a sample of preservation metadata provided by the Bibliothèque Nationale de France. The report presents the preliminary findings of Phase 1 of the Pilot and suggests that there is an opportunity to use PREMIS preservation metadata as an evidence base to support a threat assessment exercise based on the Simple Property-Oriented Threat (SPOT) model.

Digital Preservation Planning Case Study

http://www.dpconline.org/component/docman/doc_download/863-2013-may-getting-started-london-planning-case-study-ed-fay

Presentation on getting started with digital preservation planning, including scoping, risk assessing and prioritising your collection (including legacy media examples), and staff roles and responsibilities. 2013 (20 pages).

References

Consultative Committee for Space Data Systems, 2012. *Reference model for an open archival information system (OAIS): Recommended practice* (CCSDS 650.0-M-2: Magenta Book), CCSDS, Washington, DC. Available: http://public.ccsds.org/publications/archive/650x0m2.pdf

(Note this is a free to download version of ISO 14721:2012, Space Data and Information Transfer Systems – Open Archival Information System (OAIS) – Reference Model, 2nd edn).

Lavoie, B., 2014. *The Open Archival Information System (OAIS) Reference Model: Introductory Guide (2nd Edition)* DPC Technology Watch Report 14-02 October 2014. Available: http://dx.doi.org/10.7207/TWR14-02

Preservation Action

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Introduction

We know by now that digital preservation is comprised of a series of challenges emanating from organisational, resourcing, managerial, cultural, and technical issues. This section of the Handbook will focus specifically on actions that can be taken to help mitigate the technical challenges of preserving digital materials over time.

Technological obsolescence of formats

Technological obsolescence has long been considered a significant challenge of long term digital preservation. However in recent years studies have suggested that format obsolescence isn't always as prevalent as previously feared (Rosenthal 2015a, Jackson 2012). It is one issue that must be recognised and countered if digital materials are to survive over generations of technological change but It is certainly not the only challenge. 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 rather than obsolete.

A simple definition of obsolescence is the process of becoming outdated or no longer used. When talking about technological obsolescence, we refer for example to 'this Wordperfect 3.1 software is obsolete' or 'this BBC Micro computer is obsolete'. The exact moment at which obsolescence occurs can be difficult to pinpoint, particularly for materials that have only recently become obsolete. For example, just because the original application (e.g. MS Word) no longer supports a given format, it doesn't mean no other software that can read the format is unavailable. Similarly one institution may continue to use and maintain a piece of legacy software long after others have upgraded to new versions. It is perhaps therefore more useful to talk about 'institutional obsolescence', namely that the technology in question is no longer in use or easily accessed by a particular institution.

Obsolescence is an issue because all files have their own hardware and software dependencies. This was particularly the case in the early days of computing.

Change becomes an issue when it compromises the meaning of the content or its interpretation by a user. A core goal of digital preservation actions is to preserve the integrity and authenticity of the material being preserved, despite these generational changes in computing technology. In the next section we will discuss some common strategies to help minimise these changes.

Preservation strategies

In this section we review the technical strategies that can be employed to preserve digital information. After a flurry of activity in the late 1990s there has been relatively little progress in finding new strategies, though there has been significant research and development into varying implementation options and supporting technologies such as quality assurance, digital forensics (see Digital forensics), and technical representation information registries (see Technical solutions and tools in the Handbook). The techniques we will cover here are:

- Format Migration
- Emulation
- Computer Museums

Format migration

Format Migration is one of the most widely utilised preservation strategies employed and most digital preservation systems contain functionality or system data that assumes a migration solution. Format migration is different from storage media migration. It involves transferring or transforming (i.e. migrating) data from an ageing/obsolete format to a new format, possibly using new applications systems at each stage to interpret information. Moving from one version of a format standard to a later standard is a version of this method; for example moving from MS Word Version 6 (from 1993) to MS Word for Windows 2010. For frameworks and tools that are helpful for evaluating technical obsolescence of file formats see File formats and standards.

Format migration, like any intervention that has the potential to change the structure and content of data, can introduce errors and loss of information. Therefore, it is important to define metrics to measure possible loss of information and use these to do tests on the correctness and quality of format migration.

Recent work touching on quality assurance and digital preservation actions includes the work of the AQUA, SPRUCE, and SCAPE projects. To measure error rates, it is necessary to determine some very specific metrics. You might need to define what you count as an error and whether you weight some errors as being more important than others. This depends on the context/content of the record and what characteristics of the material are deemed 'significant' to preserve, as well as the migration tools and successive formats used in any migration pathway.

Some practical issues involved in this process include when to migrate – is it better to migrate from generation to generation, or should some generations be skipped? You will need to keep a record of all transformations, their results and to document detected losses of information so as to maintain evidence of authenticity and authority. PREMIS can be a useful tool for this - see the Handbook section on Metadata and documentation for more information about this standard. It is good practice always to retain the original file format as deposited to return to if required.

Emulation

Emulation offers an alternative solution to migration that allows archives to preserve and deliver access to users directly from original files. This technique attempts to preserve the original behaviours and the look and feel of applications, as well as informational content. It is based on the view that only the original programme is the authority on the format and this is particularly useful for complex objects with multiple interdependencies, such as games or interactive apps.

An emulator, as the name implies, is a programme that runs on a current computer architecture but provides the same facilities and behaviour as an earlier one. This approach has been endorsed by a number of heritage organisations, often in collaboration with technical experts and in recent years there has been some notable success in implementing emulation solutions for cultural heritage (see Resources below) . However some significant challenges remain, not least there are often rights issues associated with software licensing that need to be resolved (Rosenthal 2015b).

A particular benefit of emulation is that a single solution can be deployed to provide access to a large number of objects, so long as all those objects require delivery on the same operating system or hardware stack. Use of legacy computing equipment may however prove difficult for users, though they will almost certainly be accessing an 'authentic' representation of the records. Of course emulators have to be built and maintained, requiring a pool of expertise to be available and this cannot always be assumed. New emulators will be needed as computer architectures become obsolete, and both of these present costs and resource needs.

Computer museums

This methodology proposes the keeping of computers and their systems software (operating systems, drivers, etc.) as well as the data and applications programmes. Effort must be expended to keep all platforms in good order, and to retain all the knowledge necessary to maintain and use the machines and their programmes. The idea relies on having a source of spare parts too, but these will dwindle, as will pools of expertise. Hence this strategy tends to be an interim measure rather than a long-term solution. Some formal museums do exist, such as the Computer History Museum in California and the Centre for Computing History in Cambridge. These typically maintain machines in working order though do not provide preservation services. See also the Legacy media section of the Handbook for further information on historic file formats and media.

Implementation

The DPC Technology Watch Reports are a particularly useful guide to most common genres and file formats (including email, social media, Audio-Visual, eBooks, e-Journals, GIS, CAD, web archiving etc.) and show which strategies tend to be used most commonly in each of these areas. Tools to assist with implementation of preservation strategies are discussed in the Technical solutions and tools area of the Handbook particularly in File formats and standards.

Resources

DPC Technology Watch Report series

http://www.dpconline.org/publications/technology-watch-reports

The DPC Technology Watch Report series is intended as an advanced introduction to specific issues for those charged with establishing or running services for long term access. They identify and track developments in IT, standards and tools which are critical to digital preservation activities. They are commissioned by experts on these developments and are thoroughly scrutinised by peers before being released.

Emulation & Virtualization as Preservation Strategies

https://mellon.org/media/filer_public/0c/3e/0c3eee7d-4166-4ba6-a767-6b42e6a1c2a7/rosenthal-emulation-2015.pdf

This 2015 report on Emulation and Virtualization as Preservation Strategies by David Rosenthal was funded by the Mellon Foundation, the Sloan Foundation and IMLS. It concludes recent developments in emulation frameworks make it possible to deliver emulations to readers via the Web in ways that make them appear as normal components of Web pages. This removes what was the major barrier to deployment of emulation as a preservation strategy. Barriers remain, the two most important are that the tools for creating preserved system images are inadequate, and that the legal basis for delivering emulations is unclear, and where it is clear it is highly restrictive. Both of these raise the cost of building and providing access to a substantial, well-curated collection of emulated digital artefacts beyond reach. If these barriers can be addressed, emulation will play a much greater role in digital preservation in the coming years. (37 pages).

Systematic planning for digital preservation: evaluating potential strategies and building preservation plans

http://www.ifs.tuwien.ac.at/~strodl/paper/becker-ijdl2009.pdf

This article published in 2009 describes a systematic approach for evaluating potential alternatives for preservation actions and building thoroughly defined, accountable preservation plans for keeping digital content alive over time. The work was undertaken as part of the Europran Union-funded PLANETS project . (25 pages).

File format conversion

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

Format conversion may can help you maintain access and use of your information and mitigate risks that arise from obsolescence. This 2011 guidance from The National Archives gives you the steps you should go through in performing a file format conversion process. (29 pages).

What organizations are preserving software

http://qanda.digipres.org/1068/what-organizations-are-preserving-software

This post and responses from August 2015 on the Digital Preservation Q&A site provides a useful list and links for institutions preserving software for emulation strategies.

SCAPE Project Final best practice guidelines and recommendations

http://scape-project.eu/wp-content/uploads/2014/02/SCAPE_D20.6_KB_V1.0.pdf

This SCAPE project report published in 2014 covers three major areas: implementation of large-scale migration as a preservation strategy. Other areas are preservation of research data; and Bit preservation. (127 pages).

Case studies

The Internet Arcade

https://archive.org/details/internetarcade

The Internet Arcade is a web-based library of arcade (coin-operated) video games from the 1970s through to the 1990s from the Internet Archive, implemented using an in-browser emulation solution to provide access to the collection.

Interject

http://www.webarchive.org.uk/interject/

This prototype from the British Library demonstrates how a mixture of preservation actions can be smoothly integrated into the search infrastructure of the UK Web Archive, by acting as an 'access helper' for end users. This web page picks a few specific examples of difficult or interesting cases, and allows you to inspect what the system knows about those formats. You can also view transformed versions of those resources (combining format conversion and emulation techniques).

Rhizome

http://rhizome.org/editorial/2015/apr/17/theresa-duncan-cd-roms-are-now-playable-online/

In the 1990s, Theresa Duncan and collaborators made three videogames that exemplified interactive storytelling at its very best. Two decades later, the works (like most CD-ROMs) have fallen into obscurity. This online exhibition, co-presented by Rhizome and the New Museum brings them back, making them playable online via emulation.

Assessing Migration Risk for Scientific Data Formats

http://www.ijdc.net/index.php/ijdc/article/view/202/271

This paper explore a simple hypothesis – that, where migration paths exist, the majority of scientific data files can be safely migrated leaving only a few that must be handled more carefully – in the context of several scientific data formats that are or were widely used. The approach is to gather information about potential migration mismatches and, using custom tools, evaluate a large collection of data files for the incidence of these risks. The results support the initial hypothesis, though with some caveats.

Portico - Preservation Step-by-Step

http://www.portico.org/digital-preservation/services/preservation-approach/preservation-step-by-step

A useful step by step guide to the preservation planning and migration strategies employed by Portico The preservation plan may include an initial migration of the packaging or files in specific formats (for example, Portico migrates publisher specific e-journal article XML to the NLM archival standard).

Trash to treasure: Retro computer, software collection helps National Library access digital pieces

http://www.abc.net.au/news/2015-06-20/collecting-retro-computer-technology-to-save-digital-treasures/6560494

The National Library of Australia made public its own efforts to develop a collection of legacy computing hardware and software. It uses it to support data recovery and then implements other preservation strategies and does not rely on the computer museum for long-term preservation.

References

David Rosenthal, 2015a. *"The Prostate Cancer of Preservation" Re-examined.* Available: http://blog.dshr.org/2015/09/the-prostate-cancer-of-preservation-re.html

David Rosenthal, 2015b. *Emulation & Virtualization as Preservation Strategies*. Available: https://mellon.org/media/filer_public/0c/3e/0c3eee7d-4166-4ba6-a767-6b42e6a1c2a7/rosenthal-emulation-2015.pdf

Andrew N. Jackson, 2012. *Formats over Time: Exploring UK Web History*,. Available: http://arxiv.org/abs/1210.1714

Access

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Introduction

There has always been a strong link between preservation and access. The major objective of preserving the information content of traditional resources is so that they can remain accessible for both current and future generations. Preserving access to digital objects is the key objective of digital preservation programmes but requires more active management throughout the lifecycle of the resource before it can be assured. It is, therefore, essential to consider issues important to access provision from the beginning of the preservation process, ideally as early as the acquisition phase. This is represented within the Decision Tree for Selection of Digital Material for Long-Term

Retention included within the Acquisition and appraisal section of this handbook. With this in mind this section aims to identify the main issues that must be considered, the decisions that should be made when planning for access provision and how these may impact on preservation more generally.

Understanding users

Understanding potential users is essential when planning for the provision of access to digital objects as well as being a key consideration of broader digital preservation activities. The importance of such work is perhaps most evident in the focus on 'Designated Communities' within the Open Archival Information System Reference Model. Knowledge gained about these potential users will inform decisions made throughout the lifecycle but will likely hold most weight when choosing suitable access delivery solutions, balanced with resource and technological considerations. It is important to approach the identification of user communities and their needs systematically and objectively. In short, understanding what users want to do and what functionality can be provided by the repository.

The methodology used for the gathering of this information will vary depending on the organisational context. Potential options and tools may include the following:

- Analysis of current usage (access requests for both physical and digital objects, website statistics etc.)
- Surveys
- Focus groups
- Interviews
- Use cases
- Task analysis

When carrying out user analysis it is important to consider both existing users and non-users. Although interaction with non-users is inherently more difficult this can be a useful process towards understanding current barriers to use as well as identifying potential new market sectors.

Once collected this information should be used to inform decisions that are made in relation to the implementation of access delivery solutions. It is also important to continue to monitor the development of user communities and this should be incorporated in the standard Preservation planning activities within your organisation.

Access formats

A key consideration when planning for access is the format in which the digital objects will be delivered to the users. While there is a strong link between preservation and access in terms of the overriding objective of a digital preservation programme, there is also a need to make a clear distinction between them. There may be a combination of technical, legal, and pragmatic reasons to separate the access copy from the preservation copy, so it may be desirable or even necessary, to deliver an access copy of the digital object to the user in a different format from that held within the preservation system's storage. Indeed, an organisation may wish to offer different 'flavours' of format depending on the needs of the particular user or community in question. When selecting formats for access there are several questions an organisation will need to consider, these may include the following:

- What is the mostly commonly used/widely supported format for the object type?
- Will users have access to free viewers/software that support the proposed file type?
- What file size is produced and what are the implications for delivery to the user?
- Is the format easy to use?
- Will users require guidance or supporting documentation to allow them to access/use the objects?
- Does the organisation have separate user communities with different requirements for access?

See also File formats and standards for details of common preservation and access formats.

Legal issues for access

There are a variety of different legal issues that will probably need to be addressed when providing access to digital objects that will affect both the technological solutions that are deployed as well as who can access the material and when. This is one of the main access considerations that overlaps with acquisitions, as mentioned above, and it is essential that the correct information is gathered at that time to facilitate access requirements later in the life cycle. Without this information it may not be possible to properly manage access and may open the organisation to a number of potential legal risks.

Legal issues to be considered will include:

- Restrictions of use relating sensitivity and data protection
- Agreed embargoes on content where early access may represent a breach of contract
- Management of intellectual property rights, e.g. copyright

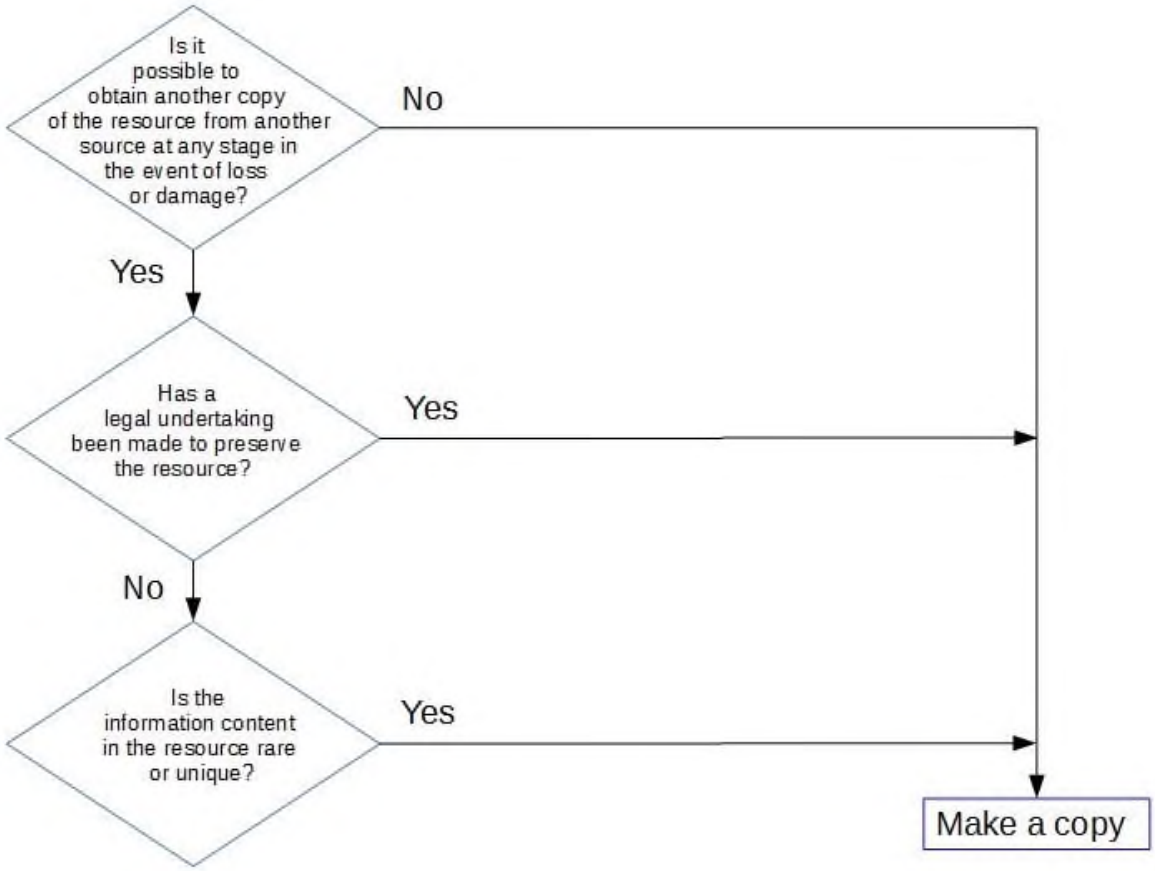
Management of IPR, in particular, should be aligned with the acquisition process with careful consideration given to transfer and ownership agreements and copyright licences put in place at that time. Licences must clearly state permitted access and reuse permissions, including third party licensing. These must then be clearly represented in policy and procedures for access, whether managed through a rights management system or by other methods.

Forms of access provision

The final key decisions an organisation must make are in the form of:

- Policy
- Procedure
- Free or charged services
- Online/Offline access, and the access environment provided
- Access for the disabled
- Storage and security

If the access copy is the only copy of a digital resource, then the danger of loss from theft or damage is clearly very high. If this approach is taken a risk assessment needs to be undertaken consisting of some of the following questions (See also Acquisition and appraisal and Storage):



Conclusions

Access is closely linked to many other digital preservation issues and technologies covered in the Handbook. In particular you may wish to look at Institutional policies and strategies, Legal compliance, Metadata and documentation, Acquisition and appraisal, Storage, Legacy media, File formats and standards, and Information security.

Resources

Born-Digital Access in Archival Repositories: Mapping the Current Landscape, Preliminary Report August 2015

https://docs.google.com/document/d/15v3Z6fFNydrXcGfGWXA4xzyWlivirfUXhHoqgVDBtUg/preview?sle=true#

This interesting document represents preliminary findings and analysis of a study and survey on current born-digital access practices in over 200 cultural heritage institutions. Respondents were primarily from the USA.

Reference model for an open archival information system (OAIS): Recommended practice (CCSDS 650.0-M-2: Magenta Book), Consultative Committee for Space Data Systems 2012

http://public.ccsds.org/publications/archive/650x0m2.pdf

This was later published as ISO 14721:2012, Space Data and Information Transfer Systems – Open Archival Information System (OAIS) – Reference Model, 2nd edition. The Access function within OAIS manages the processes and services by which consumers – and especially the Designated Community – locate, request, and receive delivery of items residing in the OAIS's archival storage. As such, it is the primary mechanism by which the archive meets its responsibility to make its information available to its user community. (135 pages).

Adrian Brown 2013 Practical Digital Preservation a how-to guide for organizations of any size

Chapter 9 (28 pages) of this book is devoted to the topic of providing access to users.

Community Owned digital Preservation Tool Registry COPTR

http://www.digipres.org/tools/

There are a large number of tools for access or that have access functionality incorporated in them. The Handbook recommends searching for them via the POWRR Grid tool within COPTR. The POWRR Tool Grid provides a set of interactive views designed to help practitioners identify and select tools that they need to solve digital preservation challenges. The Access, Use and Reuse column of the Grid identifies access tools for specific types of content or generic tools and systems that have access functions. Everything in the Grid is hyperlinked, so simply click through the displays until you find the information you are looking for. Clicking on the name of a specific preservation tool will reveal more detail on the COPTR wiki, which is where you should go to expand or update the tools information.

AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship. University of Hull, Stanford University, University of Virginia, and Yale University (2012)

http://dcs.library.virginia.edu/files/2013/02/AIMS_final.pdf

The AIMS (An Inter-Institutional Model for Stewardship) framework is a methodology for stewarding born-digital materials. It is divided into four main sections for high-level best practices for born-digital workflows: collection development, accessioning, arrangement and description, and discovery and access. Access primarily focuses on redaction and sensitive information. The appendices include, for example, sample processing workflow diagrams, an analysis of tools, and donor surveys. (195 pages).

Case studies

TNA case studies: Online access

http://www.nationalarchives.gov.uk/archives-sector/online-access.htm

a series of nine case studies published by TNA on how collections have been made more accessible by putting records online. They are drawn from a wide variety of archives.

Codebreakers: makers of modern genetics

http://www.digirati.co.uk/case-studies/wellcome/

A case study by digirati, the developers of the Wellcome Trust Library player focussing on the player its use in accessing the Francis Crick collection.The Wellcome Library's digital player is freely available for anyone to download and use. The player can be used to display all types of digital content, including cover-to-cover books, archives, works of art, videos and audio files. The software can be downloaded from the Wellcome Library GitHub account (https://github.com/wellcomelibrary/player).

Managing Risk with a Virtual Reading Room: Two Born Digital Projects, Michelle Light

http://digitalscholarship.unlv.edu/cgi/viewcontent.cgi?article=1462&context=lib_articles

Between 2010 and 2013, the University of California, Irvine, launched a site to provide online access to the personal papers of Richard Rorty and Mark Poster in the form of a virtual reading room. The virtual reading room mitigated the risks involved in providing this kind of access to personal, archival materials with privacy and copyright issues by limiting the number of qualified users and by limiting the discoverability of full-text content on the open web. The case study goes through each phase of research and thinking, including comparable projects happening at other institutions and lessons learned in a very open and informative way.

From Accession to Access: A Born-Digital Materials Case Study, by Cyndi Shein Journal of Western Archives Volume 5 Issue 1 (2014): 1-42

http://digitalcommons.usu.edu/cgi/viewcontent.cgi?article=1036&context=westernarchives

Between 2011 and 2013 the Getty Institutional Records and Archives made its first foray into the comprehensive ingest, arrangement, description, and delivery of unique born-digital material when it received oral history interviews generated by some of thePacific Standard Time: Art in L.A. project partners. This case study touches upon the challenges and affordances inherent to this hybrid collection of audiovisual recordings, digital mixed-media files, and analog transcripts. It describes the Archives' efforts to develop a basic processing workflow that applies the resource-management strategy commonly known as "MPLP" in a digital environment, while striving to safeguard the integrity and authenticity of the files, adhere to professional standards, and uphold fundamental archival principles. The study describes the resulting workflow and highlights a few of the inexpensive technologies that were successfully employed to automate or expedite steps in the processing of content that was transferred via easily-accessible media and consisted of current file formats.

Metadata and documentation

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Introduction

This section provides a brief novice to intermediate level overview of metadata and documentation, with a focus on the PREMIS digital preservation metadata standard. It draws on the 2nd edition of the DPC Technology Watch Report on Preservation Metadata. The report itself discussies a wider range of issues and practice in greater depth with extensive further reading and advice (Gartner and Lavoie, 2013). It is recommended to readers who need a more advanced level briefing.

Metadata is data about a digital resource that is stored in a structured form suitable for machine processing. It serves many purposes in long-term preservation, providing a record of activities that have been performed upon the digital material and a basis on which future decisions on preservation activities can be made in the future, as well as supporting discovery and use. The information contained within a metadata record often encompasses a range of topics. There is no clear line between what is preservation metadata and what is not, but ultimately the purpose of preservation metadata is to support the goals of long-term digital preservation, which are to maintain the availability, identity, persistence, renderability, understandability, and authenticity of digital objects over long periods of time.

Documentation is the information (such as software manuals, survey designs, and user guides) provided by a creator and the repository that supplements the metadata and provides enough information to enable the resource's use by others. It is often the only material providing insight into how a digital resource was created, manipulated, managed and used by its creator and it is often the key to others to make informed use of the resource.

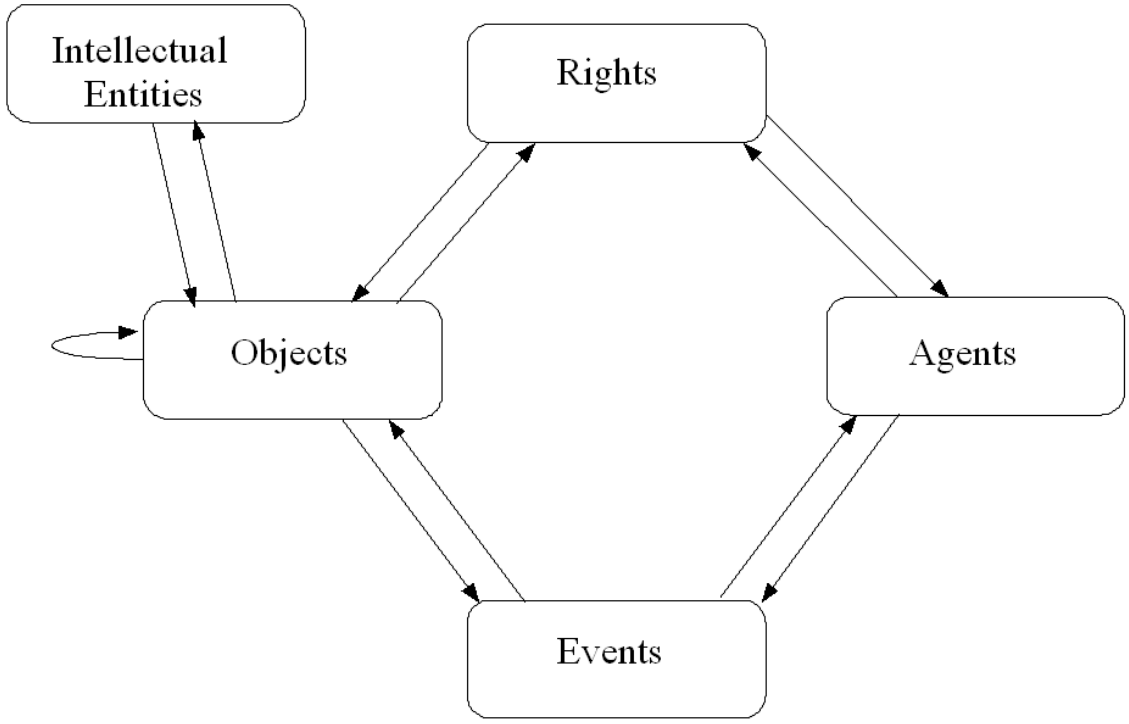
There are a number of factors which make metadata and documentation particularly critical for the continued viability of digital materials and they relate to fundamental differences between traditional and digital resources:

- **Technology**. Digital resources are dependent on hardware and software to render them intelligible. Technical requirements need to be recorded so that decisions on appropriate preservation and access strategies may be made.
- **Change**. While traditional materials may be preserved by predominantly passive preventive preservation programmes, digital materials will be subject to repeated actions, and there will be many different operators and quite possibly different institutions influencing the management of digital materials over a prolonged period of time. Recording actions taken on a resource and changes occurring as a result will provide a key to future managers and users of the resource.
- **Authenticity**. Metadata and documentation may be the major, if not the only, means of reliably establishing the authenticity of material following changes.
- **Rights management**. While traditional resources may or may not be copied as part of their preservation programme, digital resources must be copied if they are to remain accessible. Managers need to know that they have the right to copy for the purposes of preservation, what (if any) technologies have been used to control rights management and what (if any) implications there are for controlling access.
- **Future re-use**. It may not be possible for others to use the material without adequate documentation.
- **Cost**. It is expensive to create metadata manually and preservation metadata may not always be easily generated automatically. Additional metadata for digital preservation needs therefore requires careful cost/benefit trade-offs.

The PREMIS (PREservation Metadata: Implementation Strategies) Standard

PREMIS (PREservation Metadata: Implementation Strategies) is the international standard for metadata to support the preservation of digital objects and ensure their long-term usability. Developed by an international team, PREMIS is implemented in digital preservation projects around the world, and support for PREMIS is incorporated into a number of commercial and open-source digital preservation tools and systems.

The PREMIS Data Dictionary (PREMIS, 2013) is organized around a data model consisting of five entities associated with the digital preservation process:



1. **Intellectual Entity** - a coherent set of content that is described as a unit: e.g., a book
2. **Object** - a discrete unit of information in digital form, e.g., a PDF file
3. **Event** - a preservation action, e.g., ingest of the PDF file into the repository
4. **Agent** - a person, organization, or software program associated with an Event, e.g., the publisher of a PDF file
5. **Rights** - one or more permissions pertaining to an Object, e.g., permission to make copies of the PDF file for preservation purposes

Taken together, the semantic units defined in the PREMIS Data Dictionary represent the 'core' information needed to support digital preservation activities in most repository contexts. However, the concept of 'core' in regard to PREMIS is loosely defined: not all of the semantic units are considered mandatory in all situations, and some are optional in all situations. The **Data Dictionary** attempts to strike a balance between recognizing that there will be a significant overlap of metadata requirements across different repository contexts, while at the same time acknowledging that all contexts are different in some way, and therefore their respective metadata requirements will rarely be exactly the same.

Implementation

Although the PREMIS Data Dictionary is not a formal standard, in the sense of being managed by a recognized standards agency, it has achieved the status of the accepted standard for preservation metadata in the digital preservation community. A strength but also a limitation of the PREMIS Data Dictionary is that it must be tailored to meet the requirements of the specific context; it is not an off-the-shelf solution in the sense that an archive simply implements the Data Dictionary wholesale. Only a portion may be relevant in some digital preservation circumstances; alternatively, the

repository may find that additional information beyond what is defined in the Dictionary is needed to support their requirements. For example, the Data Dictionary makes no provisions for documenting information about a repository's business/policy dependencies, which may be needed to support preservation decision-making.

In short, each repository will need to invest some effort to adapt preservation metadata and documentation standards to its particular circumstances and requirements.

During implementation an institution normally identifies its own minimum standard of information required for catalogued items in the collection. Each institution can also identify its preferred levels of metadata and documentation for acquisitions and may notify and encourage suppliers or depositors to supply this information. Staff review and revise supplied information to ensure it conforms to institutional guidelines and they generate catalogue records for deposited data incorporating cataloguing and documentation standards to ensure that information about those items can be made available to users through appropriate catalogues. In many cases the contextual information for resources will be crucial to their future use and this aspect of documentation should not be overlooked.

The level of cataloguing and documentation accompanying or subsequently added to an item, and any limitations these may impose, can be documented for the benefit of future users. Where data resources are managed by third parties but made available via an institution, information may be supplied by the third party in an agreed form which conforms to institution guidelines or in the supplier's native format.

Where a need for enhanced access exists, an Institution may undertake to enhance documentation and cataloguing information to a higher standard to meet new requirements. Retrospective documentation or catalogue enhancement should also occur when the validation or audit of the documentation and cataloguing for a resource shows this to be below a minimum acceptable standard.

A significant number of both users and suppliers of preservation metadata have adopted PREMIS and many of the initial obstacles to implementation have been addressed by them. The process of implementing PREMIS in a working environment is made easier by a number of tools which can extract metadata from digital objects and output PREMIS XML. The PREMIS Maintenance Activity maintains a webpage listing the most important tools available for use with PREMIS. It also includes an active email discussion list and a wiki for sharing documents. For further information see Resources and case studies below.

See also related sections of the Handbook including Acquisition and appraisal, and Preservation planning.

Resources

PREMIS Data Dictionary for Preservation Metadata, Version 3.0

http://www.loc.gov/standards/premis/v3/index.html

The PREMIS Data Dictionary and its supporting documentation is a comprehensive, practical resource for implementing preservation metadata in digital archiving systems. The Data Dictionary is built on a data model that defines five entities: Intellectual Entities, Objects, Events, Rights, and Agents. Each semantic unit defined in the Data Dictionary is a property of one of the entities in the data model. Version 3.0 was released in June 2015 (273 pages).

Preservation Metadata (2nd edition), DPC Technology Watch Report

http://dx.doi.org/10.7207/twr13-03

This report focuses on new developments in preservation metadata made possible by the emergence of PREMIS as a de facto international standard. It focuses on key implementation topics including revisions of the Data Dictionary; community outreach; packaging (with a focus on METS), tools, PREMIS implementations in digital preservation systems, and implementation resources. Published in 2013 (36 pages).

Tools for preservation metadata implementation

http://www.loc.gov/standards/premis/tools_for_premis.php

The PREMIS Maintenance Activity maintains a webpage listing the most important tools available for use with PREMIS. This contains entries on tools, in addition to pointers to others which may be used to generate METS (Metadata Encoding and Transmission Standard - an XML schema for packaging digital object metadata) files in conjunction with PREMIS. The majority of the tools listed are for extracting technical metadata from digital objects and converting it for encoding within the PREMIS Object entity. Others can be used for checking formats, or validating files against checksums

PREMIS website

http://www.loc.gov/standards/premis/index.html

The PREMIS Editorial Committee coordinates revisions and implementation of the PREMIS standard, which consists of the Data Dictionary, an XML schema, and supporting documentation. The PREMIS Implementers' Group forum, hosted by the PREMIS Maintenance Activity, includes an active email discussion list and a wiki for sharing documents. The wiki is a particularly useful resource for new implementers, as it includes materials from PREMIS tutorials, a collection of examples of PREMIS usage and links to information on PREMIS tools. The PREMIS Maintenance Activity maintains an active registry of PREMIS implementations.

Documenting your data

http://www.data-archive.ac.uk/create-manage/document

An excellent set of resources to assist researchers with the documention and metadata for their research studies, drawn together by the UK Data Archive.

Archaeology Data Service Guidelines for Depositors

http://archaeologydataservice.ac.uk/advice/guidelinesForDepositors

The ADS Guidelines for Depositors provide guidance on how to correctly prepare data and compile metadata for deposition with ADS and describe the ways in which data can be deposited. There is also a series of shorter summary worksheets and checklists covering: data management; selection and retention; preferred file formats and metadata. Other resources for the use of potential depositors include a series of Guides to Good Practice, which complement the ADS Guidelines and provide more detailed information on specific data types.

Case studies

DPC case note: British Library ASR2 using METS to keep data and metadata together for preservation

http://www.dpconline.org/component/docman/doc_download/474-casenoteasr2.pdf

This Jisc-funded case study examines the 'Archival Sound Recordings 2' project from the British Library, noting that one of the challenges for long term access to digitised content is to ensure that descriptive information and digitised content are not separated from each other. The British Library has used a standard called METS to prevent this. July 2010 (4 pages).

Designing Metadata for Long-Term Data Preservation:DataONE Case Study

https://www.asis.org/asist2010/proceedings/proceedings/ASIST_AM10/submissions/435_Final_Submission.pdf

A short description of how PREMIS was utilized to specify the requirements for preservation metadata for DataONE (Data Observation Network for Earth) science data. 2010 (2 pages).

Preservica Case Study: Q&A with Glen McAninch, Kentucky Department for Libraries and Archives

http://preservica.com/resource/qa-glen-mcaninch-kentucky-department-libraries-archives/

Glen McAninch discusses the Importance of Provenance, Context and Metadata in Preserving Digital Archival Records.

PREMIS Implementations Registry

http://www.loc.gov/standards/premis/registry/index.php

The PREMIS Maintenance Activity maintains an active registry of over 40 PREMIS implementations with details of the repository and its use of PREMIS. Although not formally case studies, entries have details of practical experience e.g., Creating a digital repository at the Swedish National Archives using PREMIS.

References

Gartner, R. and Lavoie, B., 2013. Preservation Metadata (2nd edition), *DPC Technology Watch Report* 13-3 May 2013. Available: http://dx.doi.org/10.7207/twr13-03

PREMIS, 2013. Data Dictionary for Preservation Metadata, Version 3.0. Available: http://www.loc.gov/standards/premis/v3/index.html

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.

Reusing this information

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

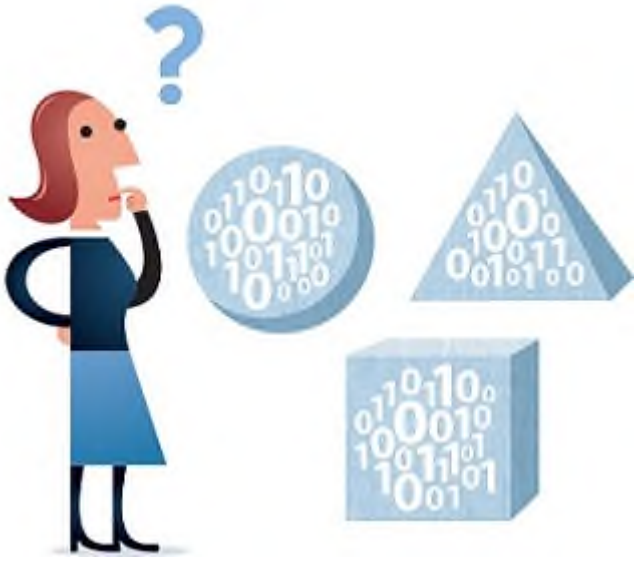


Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

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 Uses, 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.

Digital Preservation **Handbook**

Content Specific Preservation

Illustrations 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 provide a bridge to and achieve synergies with, reports in the DPC Tech Watch Series. The reports provide advanced level "deep dives" in specific areas of content preservation (e.g. email) that can be cited or to source case studies in the Handbook.
- To be developed for ease of maintenance, cost-efficiency, and sustainability in the long-term by the DPC via updates and additions to the Tech Watch series.
- To provide a brief overview and case studies, suitable for novice or intermediate level users, of digital preservation issues for specific content types covered by DPC Technology Watch Reports. Currently three content types are available: e-journals, moving picture and sound, and web-archiving. We hope to add more at a later date.

Reusing this information

Contents

e-Journals

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Overview

This case study provides a brief novice to intermediate level overview for e-journal preservation summarised from the DPC Technology Watch Report Preservation, Trust and Continuing Access for e-Journals with updates and additions by the author. Two "mini case studies" are included together with short summaries of major services and solutions. The report itself is recommended to readers who need a more advanced level briefing on the topic and practice. It covers a wider range of issues and practice in greater depth with extensive further reading and advice (Beagrie, 2013).

Introduction

Digital Preservation and trust in having continuing future access to digital content have become increasingly important for research libraries as published journals and articles have shifted from print to electronic formats. Traditional publishing business models and relationships have also undergone major transformations as a result of that shift.

Among many significant changes there has been a move from libraries purchasing and physically holding (and preserving) a paper journal locally (with multiple redundancy of copies between libraries), to renting (licensing) remote access to an electronic journal held on publishers' platforms that are often based internationally in other jurisdictions.

In parallel, there has been a growing open-access movement for e-journal articles that seeks to remove the subscription charges for access. Subscription journals, open-access journals and hybrids of the two (either a mixture of open-access and subscription articles in a journal or a 'moving wall' to open access after a fixed period of time) provide a complex landscape for the preservation of, and long-term access to, e-journals.

This e-journal landscape continues to evolve as e-publishing itself begins to shift from static to dynamic content, and the importance of data and supplementary material linked to articles increases in major disciplines.

All these changes in turn have made preservation of e-journals more demanding, more international and dependent on others, and brought issues of trust to the fore. Trust in this context is not solely of technology for preservation, but negotiating rights (and retaining a record of them for future use), and having transparent information on what is being archived, how it is preserved, and how and when it can be accessed.

This makes e-journals one of the most dynamic and challenging areas of digital preservation, particularly in terms of business models and trust mechanisms for shared or out-sourced preservation services.

Services and solutions

It is important to understand the significant implications for preservation and access of the different requirements (and terminology) that apply for e-journals: in particular the distinction between continuing access and long term preservation, as these differences lead to different types of service for e-journal archiving.

- **Continuing access** (sometimes also called post-cancellation or perpetual access) applies only to subscription journals and securing long-term access for their subscribers;
- **Long-term preservation** applies to both open and subscribed content.

The main preservation and continuing access services and solutions available for e-journals are as follows:

Keepers Registry
The Keepers Registry is a Jisc service to provide easily accessible information about inclusion of e-journals in preservation services and to highlight those e-journals for which no archiving arrangements exist. EDINA, a national data centre based at the University of Edinburgh, has developed the service along with its partner in the project, the ISSN International Centre in Paris
Legal and voluntary deposit in copyright libraries
The role of a national library is to ensure that the published heritage of its country is preserved and made accessible. In many countries legal deposit is an important vehicle for achieving this is. There is a global trend towards extending legal deposit from the print environment to cover e-journals and other electronic publications. Legal deposit legislation (or similar voluntary deposit arrangements) normally involves those subscription e-journals considered part of the national published heritage of that country. To protect the commercial interests of the publisher it also restricts off-site access to preserved electronic material for a substantial period of time. Typically this means a national legal deposit collection does not cover the international range of subscription e-journals licensed by other libraries and their users, and does not meet their requirements for 'perpetual access' rights.
CLOCKSS
CLOCKSS (Controlled LOCKSS) is a not-for-profit collaboration between libraries and publishers. It is a dark archive based on the LOCKSS software (see section below on LOCKSS) in which a limited number of libraries take on an archival role on behalf of a broader community. It provides insurance to libraries that the e-journal and other content they have subscribed to will be preserved for the long term. It is described as a 'private LOCKSS network'.

KB e-Depot
The Koninklijke Bibliotheek (KB) is the national library of the Netherlands and operates the e-Depot. It has taken the policy decision to archive journals that are within its national mandate and additionally a range of e-journals (including open-access titles in the Directory of Open-Access Journals) published beyond its borders. The e-Depot does not currently provide for post-cancellation continuing access by licensees of the content. Generally, end-user access is restricted to on-site perusal at the KB for reasons of private research only and online access is denied. However, full online access is granted to publications by open-access publishers.
LOCKSS
LOCKSS (Lots of Copies Keep Stuff Safe) provides libraries with open-source tools and support so they can take local custody of a wide variety of materials, including subscription and open-access scholarly assets (books, journals, etc.). Readers access LOCKSS preserved content whenever (and for whatever reason) the material cannot be viewed on the publisher's (or intermediary's) servers. The highly distributed nature of this approach aims to ensure that there is sufficient replication to safeguard content despite any potential disasters which might befall individual LOCKSS institutions.
Portico
Portico is designed specifically as a third-party service for scholarly literature published in electronic form and provides three specific preservation services for e-journals, e-books and digitized historical collections respectively. It provides insurance to libraries that the e-journal and other content they have subscribed to will be preserved for the long term. Portico only provides access to the e-journals they have preserved after specified 'trigger events'. In addition, if a publisher has designated Portico as such, it can also serve as a potential mechanism for post-cancellation access.
Consortial hosting
A small number of regional consortia also organize and provide their own hosting services for access and preservation of e-journals. Notable examples are OhioLink, operated by the Ohio Library and Information Network, and the Scholars Portal, operated by the Ontario Council of University Libraries.

Case study 1: the e-journal or its past issues are no longer available from the publisher

This is a highly likely scenario as publishers merge or change their business models, as larger publishers review and adjust their portfolio of titles, or as learned societies move publication contracts for their journals from one publisher to another. Journal titles are also sometimes traded between publishers, which may mean that access to past issues is no longer supported by the previous owner.

The UKSG Transfer Code of Practice initiative has produced a Code of Practice aimed at easing the problems created when journal titles move between publishers. Of relevance are the following paragraphs contained in version 3 of the code (UKSG, 2014):

The transferring publisher will alert the receiving publisher to all existing preservation arrangements for the journal.

The transferring publisher must ensure continued access to its subscribers where it has granted perpetual access rights, even if the transferring publisher will cease to host the online version of the journal after the effective transfer date. Either the transferring or the receiving publisher, or both, could fulfill perpetual access obligations. The Code intentionally does not specify the means for achieving such access, but places on the transferring publisher the responsibility for ensuring that subscribers to whom it has granted perpetual access rights will continue to have access post-transfer

The transferring publisher will use reasonable efforts to communicate journal transfer information where perpetual access rights were granted as part of a licensing agreement/Big Deal, unless archival rights will remain with the transferring publisher.

Subscribers that have been granted perpetual access rights to previously published content with the authority of the journal owner must have those rights honoured. Either the transferring or the receiving publisher, or both, could fulfil perpetual access obligations.

The receiving publisher will continue the existing, or equivalent, preservation arrangements for the journal after the effective transfer date. The receiving publisher will not remove content that was previously deposited in preserving archive(s), even if the receiving publisher will not be continuing to deposit content in the archive(s).

The decision of the publisher Sage to no longer offer its publication Graft provided a real-life example of triggered access from three archiving solutions – Portico, KB e-Depot, and CLOCKSS. In this case all were able to continue to offer access to the issues they held, either as open access (CLOCKSS and KB e-Depot) or else as a service to members (Portico). While it cannot be guaranteed that the archive will include all back issues of the title (as with Graft), participation in an archiving solution which covers at least some issues will significantly reduce the risk of disruption to continuity of service.

Case study 2: library e-Journals, perpetual access, and de-accessioning print

This case-study was first published by Jisc as part of work funded in its digital preservation programme and was incorporated into the Tech Watch Report. It has been adapted for use in the Handbook.

The case study differs from others in illustrating a few of the issues in realizing some of the potential cost savings from e-journals, particularly space savings. Increasingly, academic libraries are investing heavily in e-journals which duplicate their print back-runs. For libraries facing acute pressures on space, one solution to their problem is to dispose of or relegate print back-runs which overlap with their electronic holdings.

The case study focuses on work at Imperial College London Library in providing a database and toolkit for staff making such de-selection decisions (Cooper and Norris, 2007). Imperial established three criteria to determine the sustainability of their e-journals for de-accessioning of print. Their electronic access was classified as sustainable when at least one of the following applied:

- Imperial had perpetual access rights to the content, via the web. Imperial's perpetual access rights were nowhere near as comprehensive as they would have wished; they estimated that less than 50% of their content was covered. In addition, some of their licences specified an unsuitable delivery method for post-termination access. As they were no longer supporting networked CD-ROMs and did not have the resources to mount journal content locally, they considered a journal sustainable only if perpetual access is provided via the web.
- The journal was permanently open access for all years or certain years. Hybrid open-access journals were not included in this category, as the project was not interested in sustainability

at the article level. Finding open-access journals which fulfilled their criteria proved harder than anticipated. The main stumbling block was their need for assurance on the permanency of open access. Although the Bethesda and Berlin Declarations on Open Access include perpetual access in their definitions, Imperial discovered that not all 'open-access journals' met this criterion of permanency.

- The content was in one of Imperial's trusted services such as JSTOR, the ACM digital archive or a Jisc-funded archive. Imperial noted that of their three sustainability criteria, this one, covering services that did not offer perpetual access rights, was the hardest to pin down. The services falling into this category all shared two characteristics: the first was a good track record of stability, i.e., they had demonstrated continuity of titles from one year to another for as long as they had subscribed; the second was a history of and reputation for, affordability and value for money.

Twenty-one months into the project Imperial had identified 700 shelf-metres of sustainable stock for disposal from one site, and planned to rollout the de-selection exercise to other sites. Although it was still early days, they felt their sustainability criteria seemed to be working. The only sustainable content that they had lost was four journals from the same publisher, and they were in the process of challenging that loss. This proved to be an added benefit of the entitlements database they had created for the project; without it they would not have been aware that content over which they had perpetual access rights had been lost.

Conclusions

Continuing access and preservation of e-journals has involved initiatives in organizing multi-institutional collaboration, developing third-party services, and establishing trust in long-term access and preservation between different stakeholders. The issues it has had to address go well beyond technology, and legal, economic and service developments are equally critical to its success. Many challenges remain in e-journal archiving, but there have been significant successes and lessons learnt of interest to the wider digital preservation community as well as to libraries and publishers.

Resources

Preservation, Trust and Continuing Access for e-Journals, DPC Technology Watch Report 13-04 September 2013

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

This report 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 also includes generic lessons and recommendations on outsourcing and trust learnt in this field of interest to the wider digital preservation community. It is not solely focused on technology, and covers relevant legal, economic and service issues (43 pages).

To bin or not to bin? Deselecting print back-runs available electronically at Imperial College London Library

https://spiral.imperial.ac.uk/handle/10044/1/503

Increasingly, academic libraries are investing heavily in e-journals which duplicate their print back-runs. For libraries facing acute pressures on space, one solution to their problem is to dispose of or relegate print back-runs which overlap with their electronic holdings. This 2007 article by R Cooper and D Norris describes work at Imperial College London Library to provide a tool-kit for staff making such de-selection decisions.

UKSG, 2014 Transfer Code of Practice: Version 3.0 March 2014

http://www.uksg.org/Transfer/Code

The Transfer Code of Practice promotes a set of standards that apply whenever a journal is transferred from one publisher or publishing platform to another. Publishers who publicly sign up to the Code and apply it in practice are considered 'Transfer compliant'. As a voluntary best practices code for industry participants, the Transfer Code of Practice does not supplant contractual terms, intellectual property rights or the competitive marketplace between publishers.

CLOCKSS

http://www.clockss.org

KB e-Depot

http://www.kb.nl/en/organisation/research-expertise/long-term-usability-of-digital-resources/information-for-international-publishers

LOCKSS

http://www.lockss.org

Portico

http://www.portico.org

Ohio Link

http://www.ohiolink.edu

Scholars Portal

http://www.ocul.on.ca/node/135

Keepers Registry

http://thekeepers.org

References

Beagrie, N., 2013. Preservation, Trust and Continuing Access for e-Journals *DPC Technology Watch Report* 13-04 September 2013. Available: http://dx.doi.org/10.7207/twr13-04

Cooper, R. and Norris, D., 2007. *To bin or not to bin? Deselecting print back-runs available electronically at Imperial College London Library*, Serials 20 (3), 208–214. Available: https://spiral.imperial.ac.uk/handle/10044/1/503

UKSG, 2014. *Transfer Code of Practice*: Version 3.0 March 2014. Available: http://www.uksg.org/Transfer/Code

Moving pictures and sound

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Overview

This case study provides a brief novice to intermediate level overview summarised from the DPC Technology Watch Report on Preserving Moving Picture and Sound. Five "mini case studies" of UK collections that have run preservation and access projects for sound and moving image content are included. The report itself provides a "deep dive" discussing a wider range of issues and practice in greater depth with extensive further reading and advice (Wright, 2012). It is recommended to readers who need a more advanced level briefing on the topic and practice.

Introduction

The audiovisual domain is unique in that digitization is routinely critical to preservation. Audiovisual digitization for preservation is so pervasive that the two words have come to be used interchangeably. Audio and video need digitization for the very survival of their content, owing to the obsolescence of playback equipment and decay and damage of physical items, whether analogue or digital. The basic technology issue for collections of moving images and sound is the necessity to digitize all content currently sitting on shelves. Film on shelves can be conserved (unless it is already deteriorating), but still needs digitization to provide access.

A vital issue in preservation is access: motivation and funding for digitization purely for preservation purposes is difficult, if not impossible. There is great public, institutional and educational interest in the audiovisual record of the twentieth century. Creating access to that record is the key to obtaining the support needed for the digitization and preservation of the content.

The landscape for 'moving pictures and sound' is complicated: physically, there are large differences between audio, video and film recordings. The formats and record/playback equipment are completely separate; the digitization procedures are different; the digital files have different wrapper formats and metadata (with some overlaps); and the storage requirements differ, with video taking roughly 100 times as much storage per second of material as does audio, and high resolution digital film taking roughly 10 times more storage than video.

In addition culturally and economically, there are significant preservation and curation differences between collections from:

- **commercial media industries** – music, cinema and commercial broadcasting where preservation needs a commercial justification, a business case;
- **public bodies** – public service broadcasting, academic collections and heritage institutions such as national museums, libraries and film institutes where preservation needs a cultural heritage justification, though increasingly this sector also needs a business case;
- **technical areas** such as medicine, geology and surveillance, where recordings of images or of seismic events are raw data, kept as medical records or for reprocessing; and
- **other** – a wide range of independent collections, ranging from individual efforts to material gathered by non-profit specialist institutions (for example, steam engine clubs or ethnological research) that do not fall into any of the above categories, though their material may eventually end up being donated to a public collection.

Within the landscape is a range of technologies including engineering, computing, Internet technology, archiving, media management, museum collections management, curation, preservation, access, knowledge management and resource discovery.

Technical challenges

Audiovisual recordings are surrogate reality. The technology allows the listener and viewer to get a sensation of what a situation sounded and looked like, but the technology actually only captures the sequence of light patterns or sound pressures acting on the recording instrument (camera, microphone). These patterns (for film) and signals (for video and audio) are more like data than like artefacts. The preservation requirement is not to keep the original recording media, but to keep the data, the information, recovered from that media.

A key technology issue is moving digital content from carriers (such as CD and DVD, digital videotape, DAT and minidisc) into files. This digital to digital 'ripping' of content is an area of digital preservation unique to the audiovisual world, and has unsolved problems of control of errors in the ripping and transfer process.

The final technology area is digital preservation of the content within the files that result from digitization or ripping, and the files that are born digital. While much of this preservation has problems and solutions in common with other content, there is a specific problem of preserving the quality of the digitized signal that is again unique to audiovisual content. Managing quality through cycles of lossy encoding, decoding and reformatting is one major digital preservation challenge for audiovisual files. The other issue is managing embedded metadata.

For three decades for audio, and for at least two decades for video, archives have been digitizing their analogue content for preservation and access. The problem areas are:

- successful playback of the originals, in order to get an optimal signal to digitize;

- standards: what compression level, encoding method and file format to use; and
- efficiency: digitizing the existing analogue materials fast enough and economically enough to cope with the size and urgency of the problem.

Stages in sound and moving image digital preservation

For sound and moving image preservation, the following stages in the overall process need to be kept clear:

- **signal**: the audio from a microphone, the video signal coming out of a video camera. These signals have physical properties (bandwidth; dynamic range) that can be defined and measured. The quality of a recording and the success or failure of any process of copying, digitization or preservation can be reduced (in large part) to how well that process maintains these two physical properties of the original signal;
- **recording of a signal onto a carrier** (also called support, physical medium or recording format). For a century, the methods of capturing a signal were tied to the carrier of the signal: a wax cylinder, film reel or videotape. Digital technology produces recordings that are independent of carriers. Carrier independence is liberation: discs, tapes and films deteriorate or get damaged. Born digital recordings are liberated from these carrier-based problems, leading to a desire to liberate analogue recordings by digitization;
- **digitization**: analogue recordings can be played back and recorded onto a new carrier, or digitized and so released from carrier dependence. Digitization has to ensure that the digital version has the same bandwidth and dynamic range as the original, to capture the original quality; and
- **digital preservation of the digital representation of a signal**, meaning preserving the numbers, but also preserving the technology needed to decode (render) the numbers. Audiovisual content has a particular problem. The coding of the signal can be a compromise, not actually capturing the full signal, but instead losing some of it (lossy encoding) to get a more compact representation, thus reducing storage and transmission costs. Unfortunately coders/decoders (codecs) go out of use, and are replaced by newer technology. The file format holding the coded signal, the wrapper, is also subject to obsolescence. The failure and obsolescence of storage technology and the obsolescence of encode/decode methods and wrapper formats are major digital preservation problems for audiovisual content.

Access and rights

Sound and moving picture content arising from cinema, broadcasting and the commercial music industry is constrained by rights issues. Music has copyright protection for the composer and for the physical object containing a performance (so-called magnetic copyright). Cinema productions are protected, and music used in a film retains its separate protections. Broadcasting is even more complicated, as all the parties involved in a production may have rights in future exploitation subsequent to the one or two transmissions that were specified in typical contracts. These rights are seen as protection by rights holders, but are also seen as restrictions on access. The situation for a public broadcaster is particularly difficult. The public invariably feel that any production by a public broadcaster has already been paid for by them, is already publicly owned and should be available for public access. Unfortunately that understandable feeling is not the same as the legal definition governing when a work enters the public domain (usually determined by expiry dates on copyright and other rights).

Case study 1: The Open University (OU) Access to video assets project

This is an access and re-use project. The focus is to digitize (where necessary) audiovisual assets previously created by the OU, and place them in an asset management system so that current OU teaching and other activity can find and use these assets. Preservation is a by-product of the project rather than an end in itself. This project provides an important example of combining preservation of content with use of content, something of value to the institution in order to obtain a budget and deliver a benefit. The project was presented at the DPC Briefing Day 'Preserving Digital Sound and Vision'. The project digitized 1,200 videotapes and films, and placed the results in a Fedora digital repository. Also, 145,000 pages of documentation were digitized, providing the overall educational framework around the 1,200 items, giving them context and enhancing their ability to be re-used. The user interface provides granularity and time-based navigation. Overall this project is an outstanding example of best practice.

Case study 2: British Library Archival sound recordings project

This is a JISC-supported preservation and educational access project that ran (in its initial phase) from 2004 to 2006. A second phase added further material. Nearly 50,000 recordings of speech, music and sounds of 'human and natural environments' were digitized and placed online. The online catalogue is open to all and licensed UK further or higher education institutions can also listen to the audio. Anyone can listen to 2,000 of the items (or any of them by attending the British Library reading room in London). The differences in access between educational institutions and the general public reflects the overall issue of rights as the one remaining constraint on open access to audiovisual materials in public institutions.

Case Study 3: Imperial War Museum PSRE project

The Imperial War Museum has one of the UK's major film collections. It has been collecting film since its founding in 1919, beginning with footage from the Great War that led to the institution's founding. The Public Sector Research Exploitation (PSRE) fund made an award of nearly £1 million for cataloguing, digitization and online access (to the catalogue and the footage). The project ran from 2006 to 2009 and is of particular interest in that it is specifically aimed at commercial exploitation of a collection, and at sustainable business models around digitization and web access. The result is a website (http://film.iwmcollections.org.uk/) where anyone can view content in low quality; pull documents, stills and key frames into a lightbox; and fill a shopping basket to then purchase content.

Case Study 4: British University Film and Video Council Newsfilm Online project

This is another project with JISC sponsorship. For four decades to 1960 newsreels shown in cinemas were the main way for the general public to see moving images of current events. The initial project ran from 2004 to 2008. The results are available through a website which, as for the BL Archival Sound Recordings project, has full functionality for registered universities and colleges. The general public can see the full catalogue and can see a single key frame for each item. Since the original phase of the project, the content has been augmented by ITN/Reuters news covering the events from decades after the decline of newsreels. Newsreel items are short: the initial project provided 3,000 hours of content, but that represented 60,000 items. In addition, as with the Open University project, documentation was also placed online for context and to support search and retrieval: 450,000 pages of bulletin scripts.

Case Study 5: BFI and Regional Film Archives Screen Heritage UK (SHUK) project

SHUK is a large (£22.8 million) and complex project (involving 12 regional film archives in addition to the BFI). The project was complicated by changes in the structure and funding of the BFI, as well as a change of government and a raft of other issues. Nevertheless the project has produced major achievements:

- conservation, not digitization: construction of a £6-million vault for film conservation;
- digitization: film scanning and digital storage equipment for the regional film archives;
- access: online catalogues of regional film archive content, available to the general public.

SHUK launched on 5 September 2011 with a BBC BFI joint production, *The Reel History of Britain* (SHUK, 2011).

Conclusions

The basic technology issue for collections of moving images and sound is the necessity for digitization of all content that is currently sitting on shelves. Audio and video need digitization for their very survival, owing to obsolescence and decay of physical items, whether analogue or digital. Film on shelves can be conserved (unless it is already deteriorating) but needs digitization for access.

Playback for preservation-quality digitization implies the need for optimal recovery of the original quality, which requires professional equipment and experience. The major technical obstacle is that, for many physical formats, the needed equipment is largely obsolete, meaning that parts and repairs and skilled operators are in increasingly short supply. The urgent recommendation is, do not wait! Audiovisual holdings need to be documented and made part of a preservation plan.

The situation for sound heritage is clear. The digitization standards, encoding, wrapper and metadata are all agreed and well documented in IASA TC-04 (IASA, 2009). Uncompressed audio in the Broadcast Wave Format (BWF) wrapper is widely used and well supported. There is no reason for the basic encoding to ever be changed, though the BWF wrapper may eventually become obsolete. The only significant problem is the failure of some standard audio applications to handle embedded BWF metadata correctly (ARSC, 2011). All archives need to be aware of the risk of loss of embedded metadata. The situation for video is complex, but there is a PrestoSpace roadmap for guiding choices on the digitization of various legacy formats. There is advice from the PrestoCentre and from JISC Digital Media on the digital preservation of the resultant files. A big challenge is a registry of applications that work properly on embedded video metadata, where the diversity is huge. There is no single agreed wrapper, metadata standard or even encoding standard, and the change from standard definition to high definition brings a new set of applications, wrappers and encodings.

There is emerging technology that can improve audio (capture of the bias tone and consequent removal of temporal variation) and video transfers (direct digitization of the RF signal from the read head), which could be useful in those cases where current technology fails. So the recommendation is not to wait until such technology is further advanced and more widely available. If there are playback problems that cannot be resolved, the original audio or video format should be kept so that such advanced technology can be applied in the future.

Quality checking of the results of digitization remains an issue for video. There is a need for effective integration of signal processing technology with human checking in order to produce a really efficient method of quality control within a preservation factory approach. Quality checking is equally relevant to digital preservation – any changes or migrations due to digital obsolescence need to be checked for preservation of signal quality. Again, a purely manual approach does not scale (to the tens of millions of hours of audiovisual content in European collections), while purely algorithmic substitutes for 'looking and listening' have never been completely successful and remain an area where further research is needed.

Resources

Wright, R., 2012. Preserving Moving Pictures and Sound DPC Technology Watch Report 12-01 March 2012

http://dx.doi.org/10.7207/twr12-02

This report is for anyone with responsibility for collections of sound or moving image content and an interest in preservation of that content. New content is born digital, analogue audio and video need digitization to survive and film requires digitization for access. Consequently, digital preservation will be relevant over time to all these areas. The report concentrates on digitization, encoding, file formats and wrappers, use of compression, obsolescence and what to do about the particular digital preservation problems of sound and moving images (*33 pages*).

SHUK, 2011. Screen Heritage UK Marks new Era for Britain's Film Archives

http://www.bfi.org.uk/sites/bfi.org.uk/files/downloads/bfi-press-release-screen-heritage-uk-marks-a-new-era-for-britains-film-archives-2011-09-01.pdf

BFI Press release. 8 pages

IASA 2009 IASA TC-04, Guidelines on the Production and Preservation of Digital Audio Objects (IASA-TC 04 Second edition 2009) Canberra, IASA.

http://www.iasa-web.org/audio-preservation-tc04

This is the standard guide to digitization of audio, and the sections on metadata and digital storage are of value to all forms of digital media.

Casey, M. and Gordon, B., 2007. Best Practices for Audio Preservation. Bloomington, Indiana University Bloomington.

http://www.dlib.indiana.edu/projects/sounddirections/papersPresent/

Another audio resource (that also includes a range of digitization software tools) comes from the Sound Directions project of Harvard and Indiana Universities: much is also relevant to video digitization. (*160 pages*)

Digital Preservation Coalition Briefing day on Preserving Digital Sound and Vision, April 2011

http://www.dpconline.org/events/details/27-SoundAndVision?xref=26

This DPC briefing day in April 2011 provided a forum to review and debate the latest development in the preservation of digital sound and vision. Seven presentations (including the Open University) are linked from the programme and available to download.

ARSC Technical Committee, 2011. Study of Embedded Metadata Support in Audio Recording Software. Association of Recorded Sound Collections.

http://www.arsc-audio.org/pdf/ARSC_TC_MD_Study.pdf

A study of support for embedded metadata within and across a variety of audio recording software applications. The findings raise serious concerns, particularly for the archiving and preservation communities who rely on embedded metadata for interpretation and management of digital files representing preserved content into the future. (*21 pages*)

AVPreserve

http://www.avpreserve.com/

US based media and information management consulting firm. It website provides a range of resources for AV preservation.

BUFVC NewsFilm online Project

http://www.webarchive.org.uk/wayback/archive/20140614061518/http://www.jisc.ac.uk/whatwedo/programmes/digitisation/bufvc.aspx

British Film Institute

http://www.bfi.org.uk

the British Film Institute can advise on film and also on video – they hold a lot of video, and have a Curator for Television. Its remit is collection and preservation of film and television, and technical advice.

British Library Sound Archive

http://www.bl.uk/nsa

General technical advice on audio preservation is available from the British Library Sound Archive. Its remit is collection and preservation of all forms of audio, and technical advice.

Film Archives UK

http://filmarchives.org.uk

Collection and preservation of general audiovisual content of regional significance in the UK

JISC Digital Media

http://www.jiscdigitalmedia.ac.uk

Advice and training on still images, moving images and sound. This includes their InfoKits for Digital File Formats, Digitisation funding and sustainability, and High Level Digitisation Guide for Audiovisual Resources.

PrestoCentre

http://www.prestocentre.eu

Website provides audiovisual information, resources and advice. Access to most resources on the website requires a member subscription but a number are available to non-numbers.

The Preservation Guide Wiki

http://preservationguide.co.uk/RDWiki/

This audiovisual preservation guide was created for PrestoSpace and the BBC in May 2006. The site is now maintained as part of The Preservation Guide Consultancy. The wiki is in the public domain under a creative commons licence.

Sustaining Consistent Video Presentation

http://www.tate.org.uk/research/publications/sustaining-consistent-video-presentation

This technical paper addresses approaches to identifying and mitigating risks associated with sustaining the consistent presentation of digital video files. Originating from two multi-partnered research projects – Pericles and Presto4U – the paper was commissioned by Tate Research and is intended for those who are actively engaged with the preservation of digital video.

JISC 2009 - Archival Sound Recordings Showreel

https://www.youtube.com/watch?v=KPy9ZqWEHog

Engaging short video on British Library archival sound recordings project published on 22 Jun 2009. (6 mins 11 secs).

Further case studies

Podcasts in the Archives: Archiving Podcasting Content at the University of Michigan

http://files.archivists.org/pubs/CampusCaseStudies/CASE12.pdf

In this Society of American Archivists campus case study Alexis. A. Antracoli, University of Michigan, examines the challenges involved in developing best practices and workflows for archiving and preserving podcasting content. One major issue involved establishing standards of practice for ingest, storage, and access, especially the generation and storage of appropriate descriptive, technical, and

preservation metadata. Another challenge centered around developing the necessary technological infrastructure to support an Open Archives Information System (OAIS)-compliant system. 2010. (14 pages).

References

ARSC Technical Committee, 2011. *Study of Embedded Metadata Support in Audio Recording Software*. Association of Recorded Sound Collections. Available: http://www.arsc-audio.org/pdf/ARSC_TC_MD_Study.pdf

IASA, 2009. *IASA TC-04, Guidelines on the Production and Preservation of Digital Audio Objects*, IASA-TC 04 Second edition 2009, Canberra, IASA. Available: http://www.iasa-web.org/audio-preservation-tc04

SHUK, 2011. *Screen Heritage UK Marks new Era for Britain's Film Archives.* Available: http://www.bfi.org.uk/sites/bfi.org.uk/files/downloads/bfi-press-release-screen-heritage-uk-marks-a-new-era-for-britains-film-archives-2011-09-01.pdf

Wright, R., 2012. Preserving Moving Pictures and Sound *DPC Technology Watch Report* 12-01 March 2012. Available: http://dx.doi.org/10.7207/twr12-02

Web-archiving

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Overview

This case study provides a brief novice to intermediate level overview summarised from the DPC Technology Watch Report on Web-Archiving. Three "mini case studies" are included illustrate the different operational contexts, drivers, and solutions that can be implemented. The report itself provides a "deep dive" discussing a wider range of issues and practice in greater depth with extensive further reading and advice (Pennock, 2013). It is recommended to readers who need a more advanced level briefing on the topic and practice.

Introduction

The World Wide Web is a unique information resource of massive scale, used globally. Much of its content will likely have value not just to the current generation but also to future generations. Yet the lasting legacy of the web is at risk, threatened in part by the very speed at which it has become a success. Content is lost at an alarming rate, risking not just our digital cultural memory but also organizational accountability. In recognition of this, a number of cultural heritage and academic institutions, non-profit organizations and private businesses have explored the issues involved and lead or contribute to development of technical solutions for web archiving.

Services and Solutions

Business needs and available resources are fundamental considerations when selecting appropriate web archiving tools and/or services. Other related issues must also be considered: organizations considering web archiving to meet regulatory requirements must, for example, consider associated issues such as authenticity and integrity, recordkeeping and quality assurance. All organizations will need to consider the issue of selection (i.e. which websites to archive), a seemingly straightforward task which is complicated by the complex inter-relationships shared by most websites that make it difficult to set boundaries. Other issues include managing malware, minimizing duplication of resources, temporal coherence of sites and long-term preservation or sustainability of resources. International collaboration is proving to be a game-changer in developing scalable solutions to support long-term preservation and ensure collections remain reliably accessible for future generations.

The web archiving process is not a one-off action. A suite of applications is typically deployed to support different stages of the process, though they may be integrated into a single end-to-end workflow. Much of the software is available as open source, allowing institutions free access to the source code for use and/or modification at no cost.

Integrated Systems for Web-archiving

A small number of integrated systems are available for those with sufficient technical staff to install, maintain and administer a system in-house. These typically offer integrated web archiving functionality across most of the life cycle, from selection and permissions management to crawling, quality assurance, and access. Three are featured here.

PANDAS
PANDAS (PANDORA Digital Archiving System) was one of the first available integrated web archiving systems. First implemented by the National Library of Australia (NLA) in 2001, PANDAS is a web application written in Java and Perl that provides a user-friendly interface to manage the web archiving workflow. It supports selection, permissions, scheduling, harvests, quality assurance, archiving, and access. PANDAS is not open source software, though it has been used by other institutions (most notably the UK Web Archiving Consortium from 2004 to 2008). It is used by the NLA for selective web archiving, whilst the Internet Archive supports their annual snapshots of the Australian domain.
Web Curator Tool (WCT)
The Web Curator Tool is an open source workflow tool for managing the selective web archiving process, developed collaboratively by the National Library of New Zealand and the British Library with Oakleigh Consulting. It supports selection, permissions, description, harvests, and quality assurance, with a separate access interface. WCT is written in Java within a flexible architecture and is publicly available for download from SourceForge under an Apache public licence. The WCT website is the hub for the developer

community and there are active mailing lists for both users and developers. The highly modular nature of the system minimizes system dependencies.

NetarchiveSuite

NetarchiveSuite is a web archiving application written in Java for managing selective and broad domain web archiving, originally developed in 2004 by the two legal deposit libraries in Denmark (Det Kongelige Bibliotek and Statsbiblioteket). It became open source in 2007 and has received additional development input from the Bibliothèque nationale de France and the Österreichische Nationalbibliothek since 2008. It is freely available under the GNU Lesser General Public License (LGPL). The highly modular nature of the system enables flexible implementation solutions.

Third party and commercial services

Third party commercial web archiving services are increasingly used by organizations that prefer not to establish and maintain their own web archiving technical infrastructure. The reasons behind this can vary widely. Often it is not simply about the scale of the operation or the perceived complexity, but the business need and focus. Many organizations do not wish to invest in any skills or capital that is not core to their business. Others may use such a service to avoid capital investment. Moreover, organizations are increasingly moving their computing and IT operations into the cloud, or using a SAAS (Software as a Service) provider. Web archiving is no exception. From a legal and compliance perspective, third party services are sometimes preferred as they can provide not just the technology but also the skills and support required to meet business needs. This section introduces some of the third party services currently available but is of course a non-exhaustive list, and inclusion here should not be taken as recommendation.

Archive-It

Archive-It is a subscription web archiving service provided by the Internet Archive. Customers use the service to establish specific collections, for example about the London 2012 Olympics, government websites, human rights, and course reading lists. A dedicated user interface is provided for customers to select and manage seeds, set the scope of a crawl and crawl frequency, monitor crawl progress and perform quality assurance, add metadata and create landing pages for their collections. Collections are made public by default via the Archive-It website, with private collections requiring special arrangement. The access interface supports both URL and full text searching. Over 200 partners use the service, mostly from the academic or cultural heritage sectors. The cost of the service depends on the requirements of the collecting institution

Archivethe.Net

Archivethe.Net is a web-based web archiving service provided by the Internet Memory Foundation (IMF). It enables customers to manage the entire workflow via a web interface to three main modules: Administration (managing users), Collection (seed and crawl management), and Report (reports and metrics at different levels). The platform is available in both English and French. Alongside full text searching and collection of multimedia content, it also supports an automated redirection service for live sites. Automated QA tools are being developed though IMF can also provide manual quality assurance services, as well as direct collection management for institutions not wishing to use the online tool. Costs are dependent upon the requirements of the collecting institution. Collections can be made private or remain openly accessible,

in which case they may be branded as required by the collecting institutions and appear in the IMF collection. The hosting fee in such cases is absorbed by IMF.
The University of California's Curation Centre (UC3)
As part of the California Digital Library, provides a fully hosted Web Archiving Service for selective web archive collections. University of California departments and organizations are charged only for storage. Fees are levied for other groups and consortia, comprising an annual service fee plus storage costs. Collections may be made publicly available or kept private. Around 20 partner organizations have made collections available to date. Full text search is provided and presentation of the collections can be branded as required by collecting institutions.
Private companies
Private companies offer web archiving services particularly tailored to business needs. Hanzo Archives, for example, provide a commercial website archiving service to meet commercial business needs around regulatory compliance, e-discovery and records management. Hanzo Archives emphasize their ability to collect rich media sites and content that may be difficult for a standard crawler to pick up, including dynamic content from Sharepoint, and wikis from private internets, alongside public and private social media channels. (More details about the possibilities afforded by the Hanzo Archives service can be found in the Coca-Cola case study) Similarly, Reed Archives provide a commercial web archiving service for organizational regulatory compliance, litigation protection, eDiscovery and records management. This includes an 'archive-on-demand' toolset for use when browsing the web. In each case, the cost of the service is tailored to the precise requirements of the customer. Other companies and services are also available and readers are encouraged to search online for further options should such a service be of interest.

Case study 1: The UK Web Archive

The UK Web Archive (UKWA) was established in 2004 by the UK Web Archiving Consortium. It was originally a six-way partnership, led by the British Library in conjunction with the Wellcome Library, Jisc, the National Library of Wales, the National Library of Scotland and The National Archives (UK).

UKWA partners select and nominate websites using the features of the web archiving system hosted on the UK Web Archive infrastructure maintained by the British Library. The British Library works closely with a number of other institutions and individuals to select and nominate websites of interest. Selectively archived websites are revisited at regular intervals so that changes over time are captured.

The technical infrastructure underpinning the UK Web Archive is managed by the British Library. The Archive was originally established with the PANDAS software provided by the National Library of Australia, hosted by an external agency, but in 2008 the archive was moved in-house and migrated into the Web Curator Tool (WCT) system.

A customized version of the Wayback interface developed by the Internet Archive is used as the WCT front end and provides searchable access to all publicly available archived websites. Full text searching is enabled in addition to standard title and URL searches and a subject classification schema. The web archiving team at the library have recently released a number of visualization tools to aid researchers in understanding and finding content in the collection.

Special collections have been established on a broad range of topics. Many are subject based, for example the mental health and the Free Church collections. Others document the online response to a notable event in recent history, such as the UK General Elections, Queen Elizabeth II's Diamond Jubilee and the London 2012 Olympics.

Many more single sites, not associated with a given special collection, have been archived on the recommendation of subject specialists or members of the public. These are often no longer available on the live web, for example the website of UK Member of Parliament Robin Cook or Antony Gormley's One & Other public art project , acquired from Sky Arts.

Case study 2: The Internet Memory Foundation

The Internet Memory Foundation (IMF) was established in 2004 as a non-profit organization to support web archiving initiatives and develop support for web preservation in Europe. Originally known as the European Archive Foundation, it changed its name in 2010. IMF provides customers with an outsourced fully fledged web archiving solution to manage the web archiving workflow without them having to deal with operational workflow issues.

IMF collaborates closely with Internet Memory Research (IMR) to operate a part of its technical workflows for web archiving. IMR was established in 2011 as a spin off from the IMF. Both IMF and IMR are involved in research projects that support the growth and use of web archives.

IMR provides a customizable web archiving service, Archivethe.Net (AtN). AtN is a shared web-archiving platform with a web-based interface that helps institutions to easily and quickly start collecting websites including dynamic content and rich media. It can be tailored to the needs of clients, and institutions retain full control of their collection policy (ability to select sites, specify depth, gathering frequency, etc.). Quality control services can be provided on request. Most is done manually in order to meet high levels of institutional quality requirements, and IM has a dedicated QA team composed of QA assessors. IM has developed a methodology for visual comparison based on tools used for crawling and accessing data, though they are also working on improving tools and methods to deliver a higher initial crawl quality.

Partner institutions, with openly accessible collections for which the IM provides a web archiving service, include the UK National Archives and the UK Parliament.

Access to publicly available collections is provided via the IM website. IM provides a full text search facility for most of its online collections, in addition to URL-based search. Full text search results can be integrated on a third party website and collections can be branded by owners as necessary.

Following the architecture of the Web Continuity Service by The National Archives (The National Archives, 2010), IM implemented an 'automatic redirection service' to integrate web archives with the live web user experience. When navigating on the web, users are automatically redirected to the web archive if the resource requested is no longer available online. Within the web archive, the user is pointed to the most recent crawled instance of the requested resource. Once the resource is accessed, any link on the page will send the user back to the live version of the site. This service is considered to increase the life of a link, to improve users' experience, online visibility and ranking, and to reduce bounce rates.

Web archiving collections are available for public browsing from the IM website, a combination of both domain and selective collections from its own and from partner institutions.

Case study 3: The Coca-Cola web archive

The Coca-Cola Web Archive was established to capture and preserve corporate Coca-Cola websites and social media. It is part of the Coca-Cola Archive, which contains millions of both physical and digital artefacts, from papers and photographs to adverts, bottles, and promotional goods. Coca-Cola's online presence is vast, including not only several national Coca-Cola websites but also for example, the Coca-Cola Facebook page and Twitter stream, and other Coca-Cola owned brands (500 in all).The first Coca-Cola website was published in 1995.

Since 2009, Coca-Cola has collaborated with Hanzo Archives and now utilizes their commercial web archiving service. Alongside the heritage benefits of the web archive, the service also provides litigation support where part or all of the website may be called upon as evidence in court and regulatory compliance for records management applications.

The Coca-Cola web archive is a special themed web archive that contains all corporate Coca-Cola sites and other specially selected sites associated with Coca-Cola. It is intended to be as comprehensive as possible, with integrity/functionality of captured sites of prime importance. This includes social media and video, whether live-streamed or embedded (including Flash). Artefacts are preserved in their original form wherever possible, a fundamental principle for all objects in the Coca-Cola Archive.

Hanzo Archives' crawls take place quarterly and are supplemented by occasional event-based collection crawls, such as the 125th anniversary of Coca-Cola, celebrated in 2011. Hanzo's web archiving solution is a custom-built application. Web content is collected in its native format by the Hanzo Archives web crawler, which is deployed to the scale necessary for the task in hand.

Quality assurance is carried out with a two-hop systematic sample check of crawl contents that forces use of the upper-level navigation options and focuses on the technical shape of the site.

The Archive is currently accessible only to Coca-Cola employees, on a limited number of machines. Remote access is provided by Hanzo using their own access interface. Proxy-based access ensures that all content is served directly from the archive and that no 'live-site leakage' is encountered. The archive may be made publicly accessible in the future inside The World of Coca-Cola, in Altanta, Georgia, USA.

The Coca-Cola web archive collection contains over six million webpages and over 2TB of data. Prior to their collaboration with Hanzo, early attempts at archiving resulted in incomplete captures so early sites are not as complete as the company would like. The collection also contains information about many national and international events for which Cola-Cola was a sponsor, including the London 2012 Olympics and Queen Elizabeth II's Diamond Jubilee.

Conclusions

Web archiving technology has significantly matured over the past decade, as has our understanding of the issues involved. Consequently we have a broad set of tools and services which enable us to archive and preserve aspects of our online cultural memory and comply with regulatory requirements for capturing and preserving online records. The work is ongoing, for as long as the Internet continues to evolve, web archiving technology must evolve to keep pace.

Alongside technical developments, the knowledge and experience gained through practical deployment and use of web archiving tools has led to a much better understanding of best practices in web archiving, operational strategies for embedding web archiving in an organizational context, business needs and benefits, use cases, and resourcing options. Organizations wishing to embark on a web archiving initiative must be very clear about their business needs before doing so. Business needs

should be the fundamental driver behind any web archiving initiative and will significantly influence the detail of a resulting web archiving strategy and selection policy. The fact that commercial services and technologies have emerged is a sign of the maturity of web archiving as a business need, as well as a discipline.

Resources

Pennock, M., 2013. Web-Archiving, DPC Technology Watch Report 13-01 March 2013

http://dx.doi.org/10.7207/twr13-01

This report is intended for those with an interest in, or responsibility for, setting up a web archive. It introduces and discusses the key issues faced by organizations engaged in web archiving initiatives, whether they are contracting out to a third party service provider or managing the process in-house and provides a detailed overview of the main software applications and tools currently available.

ISO, 2012, ISO 28500:2009 Information and Documentation – the WARC file format

http://www.iso.org/iso/catalogue_detail.htm?csnumber=44717

The WARC (Web ARChive) format is a container format for archived websites, also known as ISO 28500:2009. It is a revision of the Internet Archive's ARC File Format used to store web crawls harvested from the World Wide Web.

ISO, 2013 ISO/TR 14873:2013 Information and Documentation – Statistics and quality issues for web archiving

http://www.iso.org/iso/catalogue_detail.htm?csnumber=55211

This technical report defines statistics, terms and quality criteria for Web archiving. It considers the needs and practices across a wide range of organisations such as libraries, archives, museums, research centres and heritage foundations.

Meyer E 2010 (a), Researcher Engagement with Web Archives: State of the Art Report, JISC

http://ie-repository.jisc.ac.uk/544/

This report summarizes the state of the art of web archiving in relationship to researchers and research needs focussing primarily on individual researchers and institutions.

Perma: Scoping and Addressing the Problem of Link and Reference Rot in Legal Citations

Zittrain, Jonathan and Albert, Kendra and Lessig, Lawrence, Perma: Scoping and Addressing the Problem of Link and Reference Rot in Legal Citations (October 1, 2013). Harvard Public Law Working Paper No. 13-42. Available at SSRN: http://ssrn.com/abstract=2329161 or http://dx.doi.org/10.2139/ssrn.2329161

http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2329161 or http://dx.doi.org/10.2139/ssrn.2329161

This article from the Perma project team documents a serious problem of reference rot: more than 70% of the URLs within the Harvard Law Review and other journals, and 50% of the URLs found within United States Supreme Court opinions, do not link to the originally cited information. It proposes a solution for authors and editors of new scholarship that involves libraries undertaking the distributed, long-term preservation of link contents.

Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot

http://dx.doi.org/10.1371/journal.pone.0115253

This large-scale study looked into approximately 600K links extracted from over 3M scholarly papers published between 1997 and 2012. Those were links to so-called web-at-large resources, i.e. not links to other scholarly papers. It found one out of five STM articles suffering from reference rot, meaning it is impossible to revisit the web context that surrounds them some time after their publication. When only considering STM articles that contain references to web resources, this fraction increases to seven out of ten.

The National Archives, 2010.Government Web Archive: Redirection Technical Guidance for Government Departments, version 4.2, The National Archives (UK)

http://www.nationalarchives.gov.uk/documents/information-management/redirection-technical-guidance-for-departments-v4.2-web-version.pdf

This guidance describes an innovative service that provides URL rewriting and redirection functionality for UK Government web pages by setting up redirection to the UK Government web archive where a requested URL does no longer exists on a departmental web site.

MEMENTO and the Time Travel Service

http://www.mementoweb.org/

Memento is a tool which allows users to see a version of a web resource as it existed at a certain point in the past. It is now used in several web archives. The Time Travel service based on Memento checks a range of servers including many web archives and tries to find a web page as it existed around the time of your choice.

Archive-It

http://www.archive-it.org/

Archivethe.Net

http://www.archivethe.net/en/

Hanzo Archives

http://www.hanzoarchives.com/

Internet Memory Foundation & Internet Memory Research

http://www.internetmemory.org/en/

Wayback

http://www.sourceforge.net/projects/archive-access/files/wayback/

Netarchive Suite

https://sbforge.org/display/NAS/NetarchiveSuite

PANDAS

http://pandora.nla.gov.au/pandas.html

Reed Archives

http://www.reedarchives.com/

UC3 Web Archiving Service

http://www.cdlib.org/services/uc3/was.html

Web Curator Tool

http://webcurator.sourceforge.net/

International Internet Preservation Consortium

http://www.netpreserve.org

The IIPC is a membership organization dedicated to improving the tools, standards and best practices of web archiving while promoting international collaboration and the broad access and use of web archives for research and cultural heritage. There are many valuable resources on the website including excellent short videos such as the example below.

Why Archive the Web?

https://www.youtube.com/watch?v=pU32rjTaMFE

A short video published on 18 Oct 2012 introducing the challenges of web-archiving and the IIPC. (2 mins 53 secs).

What is a Web Archive?

https://youtu.be/ubDHY-ynWi0

This short video explains 'Web Archiving' and why is it important that the UK Legal Deposit libraries support it. It was produced as part of the Arts and Humanities Research Council funded 'Big UK Domain Data for the Arts and Humanities' project.(2 mins 31 secs)

What do the UK Web Archive collect?

https://youtu.be/1QLMPIRwJEo

This video for users explains what they can expect to find and where they might go to access the three collections that the UK Web Archive hold. It was produced as part of the Arts and Humanities Research Council funded 'Big UK Domain Data for the Arts and Humanities' project. (2 mins 55 secs)

Further case studies

NDSA Website content case studies

The US National Digital Stewardship Alliance (NDSA) examines the value, opportunities and obstacles for selective preservation of the following specific web content types:

Science, Medicine, Mathematics, and Technology forums

http://www.digitalpreservation.gov/ndsa/working_groups/documents/ScienceForums_CaseStudy_public_v2.pdf

December 2013 (3 pages).

Science, Medicine, Mathematics, and Technology blogs

http://www.digitalpreservation.gov/ndsa/working_groups/documents/ScienceBlogs_CaseStudy_public_v2.pdf

December 2013 (3 pages).

Born-Digital Community and Hyperlocal News

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

February 2013 (3 pages).

Citizen Journalism

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

February 2013 (3 pages).

On the Development of the University of Michigan Web Archives: Archival Principles and Strategies

http://files.archivists.org/pubs/CampusCaseStudies/Case13Final.pdf

Michael Shallcross, Bentley Historical Library, University of Michigan details the strategies and procedures the University Archives and Records Program (UARP) followed to develop its collection of archived websites, and how it initiated a large-scale website preservation project as part of a broader effort to proactively capture and maintain select electronic records of the University. 2011 (29 pages).

References

Pennock, M., 2013. Web-Archiving, *DPC Technology Watch Report* 13-01 March 2013. Available: http://dx.doi.org/10.7207/twr13-01

The National Archives, 2010. *Government Web Archive: Redirection Technical Guidance for Government Departments*, version 4.2, The National Archives (UK). Available: http://www.nationalarchives.gov.uk/documents/information-management/redirection-technical-guidance-for-departments-v4.2-web-version.pdf

Glossary

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Introduction

Acronyms and Initials are a feature of any specialized discipline. In an emerging discipline, such as digital preservation, another major difficulty is the lack of a precise and definitive taxonomy of terms. Different communities use the same terms in different ways which can make effective communication problematic. The following working set of definitions and acronyms are those used throughout the Handbook and the DPC Technology Watch Reports and Website. They are intended to assist in its use as a practical tool.

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

A

Access As defined in the Handbook, access is assumed to mean continued, ongoing usability of a digital resource, retaining all qualities of authenticity, accuracy and functionality deemed to be essential for the purposes the digital material was created and/or acquired for.

ADS Archaeology Data Service. A UK based service active in digital preservation. **http://ads.ahds.ac.uk**

AIP Archival Information Package. An Information Package, consisting of the Content Information and the associated Preservation Description Information (**PDI**), which is preserved within an **OAIS** (OAIS term).

AMIA Association of Moving Image Archives, an organization active in the field of moving image archiving. **http://www.amianet.org**

ARC Container format for websites devised by the Internet Archive, superseded by **WARC**.

ASCII American Standard Code for Information Interchange, standard for electronic text. **https://en.wikipedia.org/wiki/ASCII**

Authentication A mechanism which attempts to establish the authenticity of digital materials at a particular point in time. For example, digital signatures.

Authenticity The digital material is what it purports to be. In the case of electronic records, it refers to the trustworthiness of the electronic record as a record. In the case of "born digital" and digitised materials, it refers to the fact that whatever is being cited is the same as it was when it was first created unless the accompanying metadata indicates any changes. Confidence in the authenticity of digital materials over time is particularly crucial owing to the ease with which alterations can be made.

B

Bit A bit is the basic unit of information in computing. It can have only one of two values commonly represented as either a 0 or 1. The two values can be interpreted as any two-valued attribute (yes/no, on/off, etc.).

Bit Preservation A term used to denote a very basic level of preservation of digital resource as it was submitted (literally preservation of the **bits** forming a digital resource). It may include maintaining onsite and offsite backup copies, virus checking, fixity-checking, and periodic refreshment to new storage media. Bit preservation is not **digital preservation** but it does provide one building block for the more complete set of digital preservation practices and processes that ensure the survival of digital content and also its usability, display, context and interpretation over time.

Born-Digital Digital materials which are not intended to have an analogue equivalent, either as the originating source or as a result of conversion to analogue form. This term has been used in the Handbook to differentiate them from 1) digital materials which have been created as a result of converting analogue originals; and 2) digital materials, which may have originated from a digital source but have been printed to paper, e.g. some electronic records.

BWF Broadcast WAV format, the European Broadcasting Union standard for a **WAV** file, with extra metadata. **http://www.digitalpreservation.gov/formats/fdd/fdd000003.shtml**

Byte (B) A unit of digital information that most commonly consists of eight bits. Historically, the byte was the number of bits used to encode a single character of text in a computer and for this reason it is the smallest addressable unit of memory in many computer architectures.

C

CCSDS Consultative Committee for Space Data Systems, the body responsible for the OAIS Reference Model. **http://public.ccsds.org/default.aspx**

Chain of Custody A key concept in forensics whereby the custody and provenance of digital hardware, media and files are safeguarded through, for example, the appointment of evidence custodians. The purpose of the Digital Evidence Bag (DEB) is to hold digitally, along with the evidential digital objects, provenance metadata that can be updated as required: a concept that is familiar to digital preservation practitioners.

Checksum A unique numerical signature derived from a file. Used to compare copies.

CLIR Council on Library and Information Resources. US based organization active in digital preservation. **http://www.clir.org**

CNI Coalition for Networked Information. US based organization active in digital preservation. **http://www.cni.org**

Continuing Access refers to the right of a subscriber to an electronic publication and their users to have on-going permanent access to electronic materials which have already been leased and paid for by the subscriber from a publisher. It is a term used, along with its synonyms perpetual access and post-cancellation access, in the information industry to describe the ability to retain access to electronic materials by the subscriber/licensee after the contractual licensing agreement with the publisher/licensor for those materials has ended, whatever the reason for the cessation. It may also cover as appropriate arrangements for digital preservation needed to guarantee some elements of continuing access.

COPTR Community Owned digital Preservation Tool Registry hosted by The Open Preservation Foundation. **http://coptr.digipres.org**

Crawl The act of browsing the web automatically and methodically to index or download content and other data from the web. The software to do this is often called a web crawler.

D

Dark Archive is an archive that cannot be accessed by any current users but may be accessible at future dates subject to the occurrence of specific pre-defined events ('**trigger event**'). Access to the data is either limited to a few set individuals or completely restricted to all.

DCC Digital Curation Centre. A UK based organization active in digital preservation. **http://www.dcc.ac.uk**

DDI Data Documentation Initiative. A de facto international metadata standard for describing data from the social, behavioral, and economic sciences. **http://www.icpsr.umich.edu/DDI**

Designated Community an identified group of potential consumers who should be able to understand a particular set of information from an archive. These consumers may consist of multiple communities, are designated by the archive, and may change over time (**OAIS** term).

Digital Archiving This term is used very differently within sectors. The library and archiving communities often use it interchangeably with digital preservation. Computing professionals tend to use digital archiving to mean the process of backup and ongoing maintenance as opposed to strategies for long-term digital preservation. It is this latter richer definition, as defined under digital preservation which has been used throughout this Handbook.

Digital Forensics The application of scientific technical methods and tools toward the preservation, collection, validation, identification, analysis, interpretation, documentation and presentation of digital information derived after-the-fact from digital sources.

Dim Archive provides bit preservation for the content plus digital preservation planning and actions for long-term perpetual access, and also limited current access (perhaps limited to on-site users or previous subscribers post-cancellation, etc.).

DigCurV Digital Curator Vocational Education Europe. A project funded by the European Commission to establish a curriculum framework for vocational training in digital curation. **http://www.digcurv.gla.ac.uk/**

Digital Materials A broad term encompassing digital surrogates created as a result of converting analogue materials to digital form (digitization), and "born digital" for which there has never been and is never intended to be an analogue equivalent, and digital records.

Digital Preservation Refers to the series of managed activities necessary to ensure continued access to digital materials for as long as necessary. Digital preservation is defined very broadly for the purposes of this study and refers to all of the actions required to maintain access to digital materials beyond the limits of media failure or technological and organizational change. Those materials may be records created during the day-to-day business of an organization; "born-digital" materials created for a specific purpose (e.g. teaching resources); or the products of digitization projects. This Handbook specifically excludes the potential use of digital technology to preserve the original artefacts through digitization. See also **Digitization** definition below.

- **Short-term preservation** - Access to digital materials either for a defined period of time while use is predicted but which does not extend beyond the foreseeable future and/or until it becomes inaccessible because of changes in technology.

- **Medium-term preservation** - Continued access to digital materials beyond changes in technology for a defined period of time but not indefinitely.
- **Long-term preservation** - Continued access to digital materials, or at least to the information contained in them, indefinitely.

Digital Preservation Management Workshop and Tutorial An intensive training workshop and online tutorial developed and maintained by Cornell University Library, 2003-2006; extended and maintained by ICPSR, 2007-2012; and now extended and maintained by MIT Libraries, 2012-on. **http://dpworkshop.org/**

Digital Publications "Born digital" objects which have been released for public access and either made available or distributed free of charge or for a fee. They may consist of networked publications, available over a communications network or physical format publications which are distributed on formats such as floppy or optical disks. They may also be either static or dynamic.

Digital Records See **Electronic Records**

Digital Resources See **Digital Materials**

Digitization The process of creating digital files by scanning or otherwise converting analogue materials. The resulting digital copy, or digital surrogate, would then be classed as digital material and then subject to the same broad challenges involved in preserving access to it, as "born digital" materials.

DIP Dissemination Information Package. An Information Package, derived from one or more Archival Information Packages (AIPs), and sent by Archives to the Consumer in response to a request to the **OAIS** (OAIS term).

DLF Digital Library Federation. A US based organization active in digital preservation. **http://www.diglib.org**

Documentation The information provided by a creator and the repository which provides enough information to establish provenance, history and context and to enable its use by others. See also **Metadata**.

DOI Digital Object Identifier. A technical and organizational infrastructure for the registration and use of persistent identifiers widely used in digital publications and for research data. The DOI system was created by the International DOI Foundation and was adopted as International Standard ISO 26324 in 2012. **http://www.doi.org**

DPC Digital Preservation Coalition. A UK and Ireland based organization active in digital preservation and responsible for the Digital Preservation Handbook. **http://www.dpconline.org**

DPTP Digital Preservation Training Program an intensive training course run by the University of London Computer Centre. **https://dptp.london.ac.uk/**

DRAMBORA Digital Repository Audit Methodology Based on Risk Assessment. A set of risk assessment tools developed by the Digital Curation Centre. **http://www.dcc.ac.uk/resources/repository-audit-and-assessment/drambora**

DROID A file profiling tool developed and distributed by TNA to identify file formats. Based on **PRONOM**. **http://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/file-profiling-tool-droid/**

E

Electronic Records Records created digitally in the day-to-day business of the organization and assigned formal status by the organization. They may include for example, word processing documents, emails, databases, or intranet web pages.

Emulation A means of overcoming technological obsolescence of hardware and software by developing techniques for imitating obsolete systems on future generations of computers.

Escrow A widespread legal practice of the deposit of content or software source code with a third party. Escrow takes place in a contractual relationship, formalized in an escrow agreement, between at least three parties: the provider, the customer, and the third party providing the escrow service.

F

FIAF International Federation of Film Archives, an association of the world's leading film archives. **http://www.fiafnet.org**

FIAT International Federation of Television Archives, a professional association for those engaged in the preservation and exploitation of broadcast archives. **http://fiatifta.org**

File Format A file format is a standard way that information is encoded for storage in a computer file. It tells the computer how to display, print, and process, and save the information. It is dictated by the application program which created the file, and the operating system under which it was created and stored. Some file formats are designed for very particular types of data, others can act as a container for different types. A particular file format is often indicated by a file name extension containing three or four letters that identify the format. **http://en.wikipedia.org/wiki/File_format**

Fixity Check a method for ensuring the integrity of a file and verifying it has not been altered or corrupted. During transfer, an archive may run a fixity check to ensure a transmitted file has not been altered en route. Within the archive, fixity checking is used to ensure that digital files have not been altered or corrupted. It is most often accomplished by computing checksums such as MD5, SHA1 or SHA256 for a file and comparing them to a stored value. **http://en.wikipedia.org/wiki/File_Fixity**

G

GIF Graphic Interchange Format, an image which typically uses lossy compression. **http://en.wikipedia.org/wiki/GIF**

Gigabyte (GB) A unit of digital information often used to describe data or data storage size, equates to approximately 1,000 **Megabytes** (MB).

GIS Geographical Information System, a system that processes mapping and data together.

H

HTML Hypertext Markup Language, a format used to present text and other information on the World Wide Web. Since 1996, versions of the HTML specification have been maintained by the World Wide Web Consortium (W3C). **http://en.wikipedia.org/wiki/HTML**

I

IASA International Association of Sound and Audiovisual Archives, an association for archives that preserve recorded sound and audiovisual documents. **http://www.iasa-web.org**

IIPC The International Internet Preservation Consortium. **http://www.netpreserve.org**

Information Assurance An aspect of digital security, specifically directed at ensuring that the quality of the information is demonstrably safeguarded, that it has not been tampered with or accessed inappropriately.

Ingest the process of turning a Submission Information Package (SIP) into an Archival Information Package (AIP), i.e. putting data into a digital archive (**OAIS** term).

InterPARES project International Research on Permanent Authentic Records in Electronic Systems. **http://www.interpares.org**

ISO International Organization for Standardization. **http://www.iso.org/iso/home.html**

J

JHove2 A characterization tool for digital objects. Characterization is comprised of four elements: identifying the object's format; validating that the object conforms to its format's technical norms; extracting technical metadata from the object; and assessing whether the object should be accepted into a repository, based on policies set by the curator. **https://bitbucket.org/jhove2/main/wiki/Home**

JPEG Joint Photographic Experts Group, a committee that oversees international standards for compression and processing of digital photographs. The majority of JPEG formats are **lossy**. **http://www.jpeg.org/**

JPEG 2000 a revision of the **JPEG** format which can use **lossless compression**.

K

Kilobyte (KB) A unit of digital information often used to describe data or data storage size, equates to approximately 1,000 **Bytes**

L

Life-cycle Management Records management practices have established life-cycle management for many years, for both paper and electronic records. The major implications for life-cycle management of digital resources, whatever their form or function, is the need actively to manage the resource at each stage of its life-cycle and to recognize the inter-dependencies between each stage and commence preservation activities as early as practicable. This represents a major difference with most traditional preservation, where management is largely passive until detailed conservation work is required, typically, many years after creation and rarely, if ever, involving the creator. There is an active and inter-linked life-cycle to digital resources which has prompted many to promote the term "continuum" to distinguish it from the more traditional and linear flow of the life-cycle for traditional analogue materials. We have used the term life-cycle to apply to this pro-active concept of preservation management for digital materials.

Lossless Compression A mechanism for reducing file sizes that retains all original data.

Lossy Compression A mechanism for reducing file sizes that typically discards data.

LOTAR (LOng Term Archiving and Retrieval) a digital preservation standard for 3D CAD models and product data management information developed by LOTAR International, an industrial consortium of aerospace and defense companies from the US and Europe. **http://www.lotar-international.org**

M

Megabyte (MB) A unit of digital information often used to describe data or data storage size, equates to approximately 1,000 **Kilobytes** (KB).

Metadata Information which describes significant aspects of a resource. Most discussion to date has tended to emphasize metadata for the purposes of resource discovery. The emphasis in this Handbook is on what metadata are required successfully to manage and preserve digital materials over time and which will assist in ensuring essential contextual, historical, and technical information are preserved along with the digital object. The **PREMIS** Data Dictionary for Preservation Metadata has become a key de facto standard in digital preservation.

METS Metadata Encoding and Transmission Standard, a standard for presenting **metadata** using **XML**. **http://www.loc.gov/standards/mets/**

Migration A means of overcoming technological obsolescence by transferring digital resources from one hardware/software generation to the next. The purpose of migration is to preserve the intellectual content of digital objects and to retain the ability for clients to retrieve, display, and otherwise use them in the face of constantly changing technology. Migration differs from the refreshing of storage media in that it is not always possible to make an exact digital copy or replicate original features and appearance and still maintain the compatibility of the resource with the new generation of technology.

MIME Multipurpose Internet Mail Extensions. A protocol for including non-ASCII information in email messages. Software typically include interpreters that convert MIME content to and from its native format, as necessary. **http://en.wikipedia.org/wiki/MIME**

MPEG Moving Picture Experts Group. A committee responsible for the development of international standards for compression, decompression, processing, and coded representation of moving pictures, audio and their combination. **https://mpeg.chiariglione.org/**

N

NCDD The Netherlands Coalition for Digital Preservation. **http://www.ncdd.nl/en/**

NDSA National Digital Stewardship Alliance a US based organization active in digital preservation. **http://www.digitalpreservation.gov/ndsa/**

NESTOR The German competence network for digital preservation. **http://www.langzeitarchivierung.de/Subsites/nestor/EN/Home/home_node.html/**

O

Open Archival Information System (OAIS) An Archive, consisting of an organization, which may be part of a larger organization, of people and systems, that has accepted the responsibility to preserve information and make it available for a **Designated Community**. It meets a set of responsibilities, as defined in section 4 of the OAIS standard that allows an OAIS Archive to be distinguished from other uses of the term ‘Archive’. The term ‘Open’ in OAIS is used to imply that the OAIS standards are developed in open forums, and it does not imply that access to the Archive is unrestricted. The OAIS abbreviation is also used commonly to refer to the Open Archival Information System reference model standard which defined the term. The standard is a conceptual framework describing the environment, functional components, and information objects associated with a system responsible for the long-term preservation. As a reference model, its primary purpose is to provide a common set of concepts and definitions that can assist discussion across sectors and professional groups and facilitate the specification of archives and digital preservation systems. It has a very basic set of conformance requirements that should be seen as minimalist. OAIS was first approved as ISO Standard 14721 in 2002 and a 2nd edition was published in 2012. Although produced

under the leadership of the Consultative Committee for Space Data Systems (**CCSDS**), it had major input from libraries and archives.

OPF Open Preservation Foundation, formerly the Open Planets Foundation. **http://openpreservation.org**

P

PAIMAS Space Data and Information Transfer Systems - Producer-Archive Interface - Methodology Abstract Standard. This **ISO**20652:2006 standard covers the first stages of the **ingest** process defined by **OAIS** reference model. **http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=39577**

PDF Portable Document Format, a set of formats and open standards maintained by the International Organization for Standardization for producing and sharing electronic documents originally developed by Adobe Systems. The original page description format has been elaborated over successive versions to enable the embedding of such complex objects as image, audio, and moving image files, hyperlinks, embedded XML metadata, and updatable forms. Specification for various versions and profiles of the format are now maintained by the International Standards Organization. **http://www.adobe.com/uk/products/acrobat/adobepdf.html**

PDF/A Versions of the PDF standard intended for archival use. **http://www.aiim.org/Research-and-Publications/Standards/Committees/PDFA**

PDI Preservation Description Information**.** The information which is necessary for adequate preservation of the Content Information and which can be categorized as Provenance, Reference, Fixity, Context, and Access Rights Information (**OAIS** term).

Perpetual Access see **Continuing Access**.

Petabyte (PB) A unit of digital information often used to describe data or data storage size, equates to approximately 1,000 **Terabytes** (TB).

PIN Pérennisation des Informations Numériques, the French national interest group for digital preservation. **http://pin.association-aristote.fr/doku.php**

Post-cancellation Access see **Continuing Access**.

PREMIS Preservation Metadata: Implementation Strategies. A de facto standard for digital preservation metadata. **http://www.loc.gov/standards/premis/**

PRONOM A database of 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. Used with **DROID**. **http://apps.nationalarchives.gov.uk/PRONOM/Default.aspx**

PST Personal Storage Table is a file extension for local 'personal stores' written by the program Microsoft Outlook. PST files contain email messages and calendar entries using a proprietary but open format, and they may be found on local or networked drives of email end users. Several tools can read and migrate PST files to other formats. **http://en.wikipedia.org/wiki/Personal_Storage_Table**

Q

R

Reformatting Copying information content from one storage medium to a different storage medium (media reformatting) or converting from one file format to a different file format (file re-formatting).

Refreshing Copying information content from one storage media to the same storage media.

S

Sandbox Containment A secure computing environment for running novel, unattested or experimental code or changes in code, including potentially malicious code. The environment is self-contained with tightly controlled resources and is characteristically virtual.

SGML Standard Generalized Markup Language an ISO standard for how to specify a document markup language or tag set. **http://en.wikipedia.org/wiki/Standard_Generalized_Markup_Language**

Significant properties Characteristics of digital and intellectual objects that must be preserved over time in order to ensure the continued accessibility, usability and meaning of the objects and their capacity to be accepted as (evidence of) what they purport to be. **https://www.archives.gov/files/era/acera/pdf/significant-properties.pdf**

SIP Submission Information Package. An Information Package that is delivered by the Producer to the **OAIS** for use in the construction or update of one or more Archival Information Packages (**AIP**s) and/or the associated Descriptive Information (OAIS term).

SMPTE Society of Motion Picture and Television Engineers, a professional organization and technical standards body for television and motion picture. **https://www.smpte.org**

T

TDR Trusted Digital Repository. A trusted digital repository has been defined as having “a mission to provide reliable, long-term access to managed digital resources to its designated community, now and into the future”. The TDR must include the following seven attributes: compliance with the reference model for an Open Archival Information System (**OAIS**), administrative responsibility, organizational viability, financial sustainability, technological and procedural suitability, system security, and procedural accountability. The concept has been an important one particularly in relation to certification of digital repositories.

Terabyte (TB) A unit of digital information often used to describe data or data storage size, equates to approximately 1,000 **Gigabytes** (GB).

Three-Legged Stool A conceptual approach to digital preservation that suggests a fully implemented and viable preservation program addresses organizational issues, technological concerns, and funding questions, balancing them like a three-legged stool. Developed as part of the **Digital Preservation Management Workshop and Tutorial**.

TIFF Tagged Image File Format, a common format for images typically **lossless**. **http://en.wikipedia.org/wiki/Tagged_Image_File_Format**

TRAC Trusted Repository Audit and Certification, toolkit for auditing a digital repository. **http://www.crl.edu/sites/default/files/d6/attachments/pages/trac_0.pdf**

Trigger Event This terminology is used when specific conditions relating to an electronic publication and its continued delivery to users are met. If the publication is no longer available to users from the publisher or any other source for a variety of reasons then a trigger event is said to have occurred. They can set in motion access for users via an archive where the electronic publication may be digitally preserved.

U

UKWA UK Web Archive. **http://www.webarchive.org.uk/ukwa/**

V

W

WARC The WARC (Web ARChive) format is a container format for archived websites, also known as ISO 28500:2009. It is a revision of the Internet Archive's ARC File Format used to store web crawls harvested from the World Wide Web. **http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=44717**

WAV the standard file wrapper for audio; see BWF (Broadcast WAV Format) for the professional variant. **http://en.wikipedia.org/wiki/WAV**

Writeblockers Tools that prevent an examination computer system from writing or altering a collection or subject hard drive or other digital media object. Hardware writeblockers are generally regarded as more reliable than software writeblockers.

X

XML Extensible Markup Language, a widely used standard (derived from **SGML**), for representing structured information, including documents, data, configuration, books, and transactions. It is maintained by the World Wide Web Consortium (W3C). **http://www.w3.org/XML/**

Y

Z