Управление цифровыми материалами и их долгосрочная сохранность

Digital-asset management and long-term preservation

Управління цифровими матеріалами та їх довготермінове збереження

Роман Пионтек
Компания «Ex Libris»,
Гамбург, Германия

Roman Piontek
Ex Libris,
Hamburg, Germany

Роман Піонтек
Компанія «Ex Libris»,
Гамбург, Німеччина

Главной функцией репозитория цифровой информации является хранение, управление и представление цифровых фондов библиотеки. С началом процесса оцифровки начинают цифровые фонды стремительно расти. Если речь идет о больших, а, в данном случае, об огромных масштабах оцифровки, то на первый план выходит задача обеспечения доступа к цифровым объектам. Чем больше размеры цифровой коллекции, тем более актуален для библиотеки вопрос об обеспечении ее сохранности и особенно длительной сохранности. Таким образом, библиотека должна разработать механизм, обеспечивающий длительное хранение документов, тесно связав его с процессом оцифровки и предприняв шаги для минимизации затрат, связанных с размещением управлением фондом. Для обеспечения долговременного сохранения этих цифровых изображений библиотека должна выстроить систему хранения. Необходимо сохранять сведения об источнике файла (или сам оригинал), иметь возможность отслеживать историю всех изменений, происходивших на протяжении всего времени хранения; для этого файлы хранятся в оригинальном формате TIFF, дополняющем формат JPEG. Метаданные, описывающие содержание каждого изображения, а также его технические характеристики, также должны быть сохранены как независимый объект. Кроме того, программное обеспечение такой система сохранности позволяет библиотеке отслеживать совместимость файлов с существующими версиями программного обеспечения для представления изображений и поддерживать рабочие потоки, гарантируя, в случае необходимости, безопасную миграцию файлов в релевантный формат.

The core functionality of a digital repository is storing, managing and representing the library's digital collections. Once digitization process is established, digital collections start growing rapidly. In digitization projects of bigger or, like in this case, huge scope, the accessibility to the digital objects becomes a requirement with high priority. The bigger the size of the digital collection, the aspect of preservation and especially long-term preservation becomes crucial to the library. The library must, therefore, include a mechanism to ensure long-term preservation and combine it tightly with the digitization process in order to reduce significantly hosting and maintenance efforts.
For long-term preservation of these digital images, the library would require a preservation system. In order to maintain the provenance of the image files and track the history of any changes made over time, they would be stored in the original TIFF format in addition to JPEG. Metadata describing the content of each image as well as its technical properties would be stored as a self-contained object. In addition, the preservation system enables the library to monitor the compatibility of the files with current image presentation software versions and to support workflows for safe migration of the files to a relevant format when necessary.

Головною функцією репозитарію цифрової інформації є збереження, управління та представлення цифрових фондів бібліотеки. З початком процесу оцифрування цифрові колекції бібліотеки починають стрімко зростати. Якщо йдеться про великі чи, як у даному випадку, про величезні масштаби оцифрування, то на перший план виходить завдання забезпечення доступу до цифрових об'єктів. Чим більший розмір цифрової колекції, тим актуальнішим для бібліотеки є питання про забезпечення її збереження та, особливо, довготермінового збереження. Таким чином, бібліотека повинна розробити механізм, що забезпечить довготермінове збереження документів, тісно пов'яже його з процесом оцифрування та здійснить необхідні дії для мінімізації витрат, що пов'язані із утриманням та управлінням фондом.

Для забезпечення довготермінового збереження цифрових зображень бібліотека має вибудувати систему збереження. Необхідно зберігати відомості про джерело файлу (або сам оригінал), мати можливість відслідковувати історію усіх змін, що відбуваються протягом усього часу збереження; для цього файли зберігаються в оригінальному форматі TIFF, який доповнює формат JPEG. Метадані, що описують зміст кожного зображення, а також його технічні характеристики, також повинні бути збережені як незалежний об'єкт. Окрім того, програмне забезпечення такої системи збереження дозволяє бібліотеці відслідковувати сумісність файлів з існуючими версіями програмного забезпечення для представлення зображень і підтримувати робочі процеси та, за необхідності, гарантувати безпечну міграцію файлів у релевантний формат.

1. Repository for Storing and Managing of Digital Objects / Digital Collections

The core functionality of a digital repository is the storage, management and presentation of the library's digital collections. Digital collections are growing rapidly once a digitization process has been established. In digitization projects of bigger or in this case huge scope, the accessibility to the digital objects becomes a requirement with high priority. The bigger the size of the digital collection, the aspect of preservation and especially long-term preservation becomes crucial to the library. The library must, therefore, include a mechanism to ensure long-term preservation and combine it tightly with the digitization process in order to reduce significantly hosting and maintenance efforts.

For storing, managing and presenting digital objects an interface for the submission of the objects through the digitization software to the repository has been developed. Already digitized items (JPEG, TIFF, BMP …) and expected to-be-digitized items can be ingested into the repository by using the same mechanisms. Various file formats can be stored:

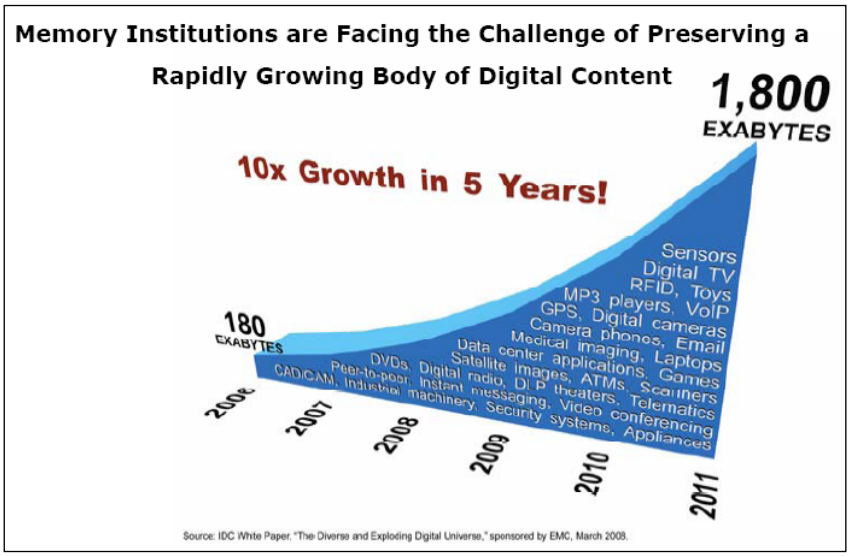
- Digitized mediaeval books
- Modern literature
- Audio material
- Video material
- Digitized artwork
- Various photographic material
- 3D graphics

The Saxonian State and University Library at Dresden (Germany) can be cited as an example. This library is using scanner technology from Zeutschel plus the software Goobi which works in conjunction with the digital asset management system Rosetta, provided by Ex Libris.

2. Long-Term Preservation

2.1. The Challenge

The need for preserving digital assets is only a few decades old, but is one that becomes more pressing and keeps growing by the day. According to the International Data Corporation (IDC), by 2011 information that is either created, captured, or replicated in digital form has increased 10 fold over that produced in 2006 (see image below). The compound annual growth rate between 2006 and 2011 is almost 60%.



In the last two decades, digital technology has enabled us to create, use, and be enriched by information in ways that were unthinkable a generation ago. With the technology available today, any internet user can become a global publisher in a matter of minutes by posting video to YouTube. Researchers can use the Web for posting scholarly communication or for documenting experiments and research.

This data may in turn be deposited into an institutional repository and searchable through a discovery platform. The very same technological advances which make obtaining and sharing information so easy also pose some modern challenges. A 4,000-year-old stone tablet displayed at the Metropolitan Museum of Art is still readable today. Without proper preservation, the digital image showing the tablet on the museum's website may not be viewable in as little as ten years from now.

Possibly of even greater concern is the preservation of the metadata related to the tablets. If the records describing the tablet – which are all stored digitally today – become unreadable, this knowledge will be lost for future generations. The problem of preserving digital material is most acute when it comes to items that have been produced in digital format that is born digitally. Representing over 93% of the world's information, the vast majority of this material exists exclusively in digital format, a fact that makes the preservation of digital information critical to the perpetuation of our cultural heritage and collective memory.

The preservation of digital information is dependent on a number of components including storage media, hardware, operating system, software application, and file format. If any of these is missing or does not function properly, the digital object may become unusable:

- The media on which digital information is stored have a very limited usable lifespan.
- The preservation of digital material is also dependent on the applicability of the hardware, operating system, or software application required to operate them. It may no longer be readily available.
- If all of the other components are in place to correctly read and display digital files, these files may yet become corrupted as a result of malicious virus attacks or hardware or software failures.

While organizations that digitize have systems in place for storing and managing digital objects, these systems were not designed with preservation in mind. Digital preservation is about risk management – guaranteeing future usability of and accessibility to digital content. To illustrate the differences, we can examine how digital scans of a special collection of rare books owned by a library are handled:

To make the collection available to patrons online, the library would store the digital images in a Digital Repository system. As the emphasis in this case is on providing users with access to easily view the images, these would likely be converted to JPEG or JPEG 2000 formats.

For long-term preservation of these digital images, the library would require a preservation system. In order to maintain the provenance of the image files and track the history of any changes made over time, they would be stored in the original TIFF format in addition to JPEG. Metadata describing the content of each image as well as its technical properties would be stored as a self-contained object. In addition, the preservation system enables the library to monitor the compatibility of the files with current image presentation software versions and to support workflows for safe migration of the files to a relevant format when necessary.

It is also worth noting that although preservation focuses on risk management, it would be a mistake to equate preservation with disaster recovery and backup, as indicated in the Digital Preservation briefing paper published by the Joint Information Systems Committee (JISC) in the UK:

Disaster recovery strategies and backup systems are not sufficient to ensure survival and access to authentic digital resources over time. A backup is a short-term data recovery solution following loss or corruption and is fundamentally different to an electronic preservation archive.

2.2. The Cost of Lost Information

Digital preservation is an expensive undertaking. A recent study by the Academy of Motion Picture Arts and Sciences provides a vivid example of the magnitude of the costs involved:

- yearly cost of saving a film master: US$ 1,059
- yearly cost of saving a digital master converted from film: US$ 12,514
- yearly cost of saving a digital master born digitally: US$ 208,569

While the cost of preservation may be high, the cost of recreating a digital resource may be much higher. For example, a NASA study has estimated the cost of preservation for electronic engineering records at about US$ 5-7 per megabyte per year; according to the same study, the cost of recreating the information is estimated at about US$ 1,250 per megabyte. According to the 2005 Heritage Health Index Report published by Heritage Preservation and The Institute of Museum and Library Services, over half of the digital materials in U.S. collections – including DVDs, CDs, and MP3 files – are in probable jeopardy.

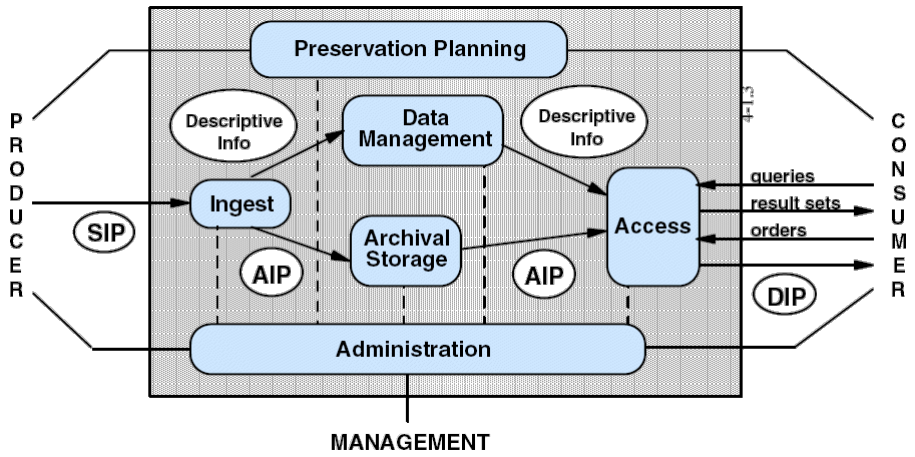
Given that many organizations do not know the extent of the problem they face, it is not surprising that the loss of digital data is commonplace, and in some circumstances seems to be accepted as an inevitable hazard. Only 29% of respondents in a survey by the Digital Preservation Coalition in the UK could say they have not lost access to some digital information, as a result of it being impossible or too expensive to recover. Even when referring to their most important type of data, this proportion only rose to 43%.

2.3. Requirements of a Digital Preservation System

The Open Archival Information System (OAIS) reference model describes the characteristics of a digital preservation system. Written by the Consultative Committee on Space Data Systems, it is now an ISO standard. The model has become widely accepted among preservation bodies and experts worldwide and has been used as a guideline to evaluate current implementations of preservation and archiving initiatives.

The OAIS model (see graphic below) describes six high level functions that must be present in a preservation system:

- Ingest
- Storage
- Data Management
- Administration
- Preservation Planning
- Access



The following table provides a high level overview of the requirements associated with digital long-term preservation and the most common limitations of simple digital repository systems in addressing these requirements. It is therefore important to evaluate systems against these issues. This is not an attempt to cover all the requirements of the OAIS model but rather to highlight some of the functions that are more relevant to preservation.

Digital Preservation Challenges	Preservation System Requirements	Limitations of Digital Repositories
Ingest		
How do we manage the flow of incoming material (drinking from the fire hose)?	A digital preservation system should be capable of ingesting large volumes of material automatically.	***Very limited:*** «…the handicraft approach will not scale to support the longevity of digital content in the diverse and large digital libraries...»[1]
Storage		
How do we manage storage capacity to handle the constant inflow of large quantities of material?	Advance storage space planning is an important function that ensures adequate capacity at all times.	***Very limited***: No special tools for storage capacity are planned as storage requirements are more predictable.
How do we store objects to ensure the long term viability of the information?	Objects and their metadata are self-contained and comply with the Preservation Metadata: Implementation Strategies (PREMIS) data model.	***Very limited:*** Objects and metadata are assumed to be rendered by external software applications and are not self-contained.
How do we protect stored information?	In addition to standard measures such as backup, virus protection, and redundant storage, objects are stored in a permanent repository, which is separate from the deposit and staging storage areas.	***Very limited:*** Digital repositories do not have a separate permanent storage area. New material is entered directly into the "permanent" storage area.
Data Management		
How do we manage metadata associated with the objects we store?	Preservation requires the ongoing management of objects and their associated metadata.	***Adequate:*** Digital repositories are generally required and equipped to perform similar data management functions.
Administration		
How do we deploy a system that is flexible enough to handle future and yet un-known requirements?	To ensure long-term viability, the administration function must allow the embedding of additional tools to accommodate the ever-changing world of technology and formats.	***Limited:*** Digital repositories are built with a shorter time horizon in mind and are not generally highly adaptable in nature.

[1] Keynote Address at the 11th European Conference on Digital Libraries (ECDL), Budapest (17 September 2007). ©Seamus Ross, HATII at the University of Glasgow

Digital Preservation Challenges	Preservation System Requirements	Limitations of Digital Repositories
Preservation Planning		
How do we ensure authenticity and integrity of the material over time?	The system must store the object's original representation as well as a detailed audit trail recording the history of its change.	***Very limited:*** Not part of the functionality of a digital repository.
How do we know when an object is at risk of being inaccessible?	The system must support a risk analysis process with integration to technical registries such as PRONOM.	***Very limited:*** Not part of the functionality of a digital repository
Access		
How do we manage ongoing accessibility to stored content while maintaining its compliance with ownership rights and user access privileges?	To protect stored material, a preservation system creates publishable copies of the material that can be used for accessing the stored content.	***Adequate:*** A Digital Repository is focused on access management to stored objects.

2.4. A Practical Approach to the Digital Preservation Challenge

Libraries, museums, and academic institutions that have long taken a role in preserving our collective knowledge and cultural heritage can no longer ignore the digital preservation challenge. Given the high cost of preservation and the even higher potential cost of a lack thereof, a practical approach to digital preservation is required. The following steps provide a direction and recommendations for organizations that are looking to take action.

1. Recognize the need for preservation and create executive accountability for implementing the process. One of the common issues with preservation is the absence of accountability since the results of preservation (or lack thereof) can only be seen in the long run.
2. Translate awareness into action by establishing a team to lead the preservation effort.
3. Develop detailed plans to conduct an information audit, a preservation assessment, and a risk analysis in order to determine the organization's preservation needs.
4. When evaluating a solution to address your preservation needs, look for a system that supports preservation as a process, not just the storage of digital objects.
5. Given the extent of the investment and effort required, look at the implementation of a preservation system as a multi-stage effort.
6. The cost of developing and maintaining a preservation system makes it prohibitive for any single organization to do alone. A collaborative approach is usually the only practical way to develop and maintain a preservation system in the long run.

A long-term preservation system capable to fulfill the requirements of a national library is, for example, the system Rosetta provided by Ex Libris. Libraries that have implemented the long-term preservation system Rosetta and could act as successful examples include Saxonian State and University Library at Dresden (Germany) and the National Library of New Zealand.