

DELIVERABLE

Project Acronym: DCA
Grant Agreement number: 270927
Project Title: Digitising Contemporary Art

D6.1 Guidelines for a Long-term Preservation Strategy for Digital Reproductions and Metadata

Revision: 1.0

Author(s): Sofie Laier Henriksen, Wiel Seuskens & Gaby Wijers (NIMk)
External reviewer: Robert Gillesse (DEN)

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
P	Public	x
C	Confidential, only for members of the consortium and the Commission Services	

REVISION HISTORY AND STATEMENT OF ORIGINALITY

Revision History

Revision	Date	Author	Organisation	Description
V0.1	09/09/2011	SLH / WS /GW	NIMk	Text
V0.2	19/09/2011	EM	IBBT	Internal review
V0.2	26/09/2011	RV	PACKED vzw	Internal review
V0.3	11/12/2011	SLH / WS / GW	NIMk	Amendments
V0.4	03/01/2011	RV	PACKED vzw	Internal review
V0.5	18/01/2012	RG	DEN	External peer review
V0.6	25/01/2012	SLH / GW	NIMk	Final amendments
V1.0	13/02/2012	RV	PACKED vzw	Final version

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

CONTENT TABLE

1. Executive summary	4
2. Introduction	5
3. Definitions of digital preservation	8
4. Trustworthy repositories and file authenticity	9
5. Standardisation model OAIS	10
6. How to get started	14
6.1 Topics of the Digital Preservation Capability Model	15
7. Bit preservation and logical preservation	17
7.1 Bit preservation.....	17
7.2 Logical preservation	17
8. Introduction to recommended file formats and containers	19
8.1 Codecs and Containers	19
8.2 Uncompressed, lossless and lossy compression	19
8.3 Recommended file format properties	21
8.4 Recommended file format standards.....	22
8.5 Presentation and Web use	26
9. Methods for preservation	27
9.1 Migration	27
Issues/Risks.....	27
Recommendations for migration procedures.....	28
9.2 Migration - Normalisation.....	28
9.3 Technological Hardware preservation	28
9.4 Emulation.....	28
The creation of an emulator.....	29
Issues/risks	29
9.5 Emulation – Virtualisation	30
9.6 Encapsulation	30
Issues/Risks.....	31
9.7 Cloud Computing.....	31
Issues/Risks.....	32
9.8 How to choose.....	32
10. Preservation planning	33
11. Preservation Metadata	36
11.1 Recommended properties of metadata	36
11.2 What is preservation metadata?.....	36
Recommendations for what to include.....	37
11.3 Recommendation of preservation metadata standards.....	39
11.4 Extraction and creation of metadata.....	40
12. Checksums	43
13. Persistent identifiers	45
Glossary	47
References	49
ANNEX I - Digital Preservation Capability Performance Metrics	54
ANNEX II - File formats and codec suitable for Web and presentation	61
ANNEX III - Manual for metadata extraction	63
ANNEX IV - Example of a metadata schema with PREMIS	64

1. Executive summary

These *Guidelines for a long-term preservation strategy for digital reproductions and metadata* explains how to preserve digital materials such as text, images and video. It gives a theoretical introduction to the subject as well as practical examples of how to manage a collection of digitised and born-digital artworks. Every institution with a digital repository should have a policy and plan to ensure access to content, not only today but also in the future. If digital preservation is not taken into account, the risk of losing or changing data will become inevitable. Preservation should therefore be a part of digital collection management.

An introduction to the Open Archival Information System (OAIS) is presented with examples of how it can be used for the Digitising Contemporary Art (DCA) project, presenting the elements of digital preservation and workflow. Digital preservation can seem overwhelming but by using the Digital Preservation Capability Model, explained in these guidelines, every collecting institution can assess how well it is doing in the different areas. It indicates where to start the process and what to focus on.

Preservation starts at the beginning of a digitisation process, when one chooses an appropriate file format or codec. The differences between codecs and containers as well as uncompressed, lossy compression and lossless compression are explained. Recommended file formats for texts, images, audio and video are listed; the selection is based on the preferred preservation properties. A recommended file format should (among others) be uncompressed, well-known, supported by several types of software and platforms and preferably an acclaimed open standard.

The file formats that a collecting institution uses do not always meet these requirements. Sometimes there is a good reason to use a more 'inappropriate' file format. This might be the case when the source is already compressed, the source is low-quality or if one doesn't have the possibility to buy or use the best equipment. When the right file formats are chosen, it is important to make sure the data is stored in the right way. This can be partially achieved by having several copies on different types of hardware at different geographical locations. This lowers the risk of any loss or change of bits. Besides bit-preservation it is vital to consider how to keep the logical understanding of the material long-term.

Making a preservation plan and choosing the right method for the logical preservation of data is important. A well-known method is called migration. It refers to the transfer of data from hardware to hardware, or to the conversion from one file format into another. Emulation or virtualisation can be considered for more complex data or issues with obsolete data like software programmes. Other aspects of preservation are the creation and storage of preservation metadata, the validation of files and the creation of checksums. These *Guidelines* recommend tools for such procedures and also include a manual for a basic programme called Exiftool.

These *Guidelines* are intended to be used by participants of the DCA project, but can also be used by other institutions in the process of digitising their collections. The different elements of digital preservation are explained in a basic and accessible way that is not too technical. Together with some examples from real life situations and recommendations for practical tools, this approach should give the basics needed for collection managers to create a suitable preservation policy and plan. It is important to keep in mind that digital preservation is a continuous process that has to be regularly re-evaluated by the collecting institution.

2. Introduction

The Digitising Contemporary Art (DCA) project aims to digitise contemporary art objects from 12 European countries and make them accessible to the wider public through Europeana – a single access point for European cultural heritage. The project partners include 21 art institutions and museums and 4 technical institutions. The project started on January 1, 2011 and shall last until July 2013. The European Commission and all partners support the project financially. More information can be found on www.dca-project.eu.

“The Digital Agenda for Europe, our ICT policy agenda through to 2020, would simply not be complete if we ignored the content dimension of our vision for the future. Internet is eager for high quality content from reliable sources.”

Neelie Kroes

The project is made up of work packages, all concerning different topics relevant for digitisation. Each work package leader is responsible for one such package, but can also be involved within others. All participating institutions contribute information to the different work packages and can use the results. Work Package 6 of the DCA project is on digital preservation and has been delegated to the Netherlands Media Art Institute (NIMk). In order to safeguard the digital materials that are accessible today, it is important for an institution with any kind of digital repository to have a policy and plan on how to preserve and access the content for future use. Otherwise one risks losing the information. This applies for born digital materials as well as for digitised analogue material. The focus is different when preserving analogue materials, where the focus is on the condition of physical carriers and the storage environment.

“There is probably no greater ambition than to perpetuate our rich cultural heritage. It is therefore in full consciousness of our responsibility towards past and future generations and in deep humility that we have approached our mission. Digital preservation is a key challenge for the digital age. Digital materials have become an integral part of our cultural and scientific heritage. Due to rapid technical change, however, digital objects are more endangered than they appear.”

**Comité des Sages (Reflection Group on Bringing Europe’s Cultural Heritage Online)
- ‘The New Renaissance’**

This paper holds a guideline and basic introduction to relevant strategies, planning tools, file formats, codecs and metadata standards needed for making a well-founded preservation plan for digital material. It provides responses to the main problems encountered by choosing specific strategies, and guidelines to help identify problems and potential solutions associated with digital materials of different kinds. The focus is on making sure that the data is suitable for long-term storage and prepared in such a way that it can still be accessed and used in the future. It does not take into account how data can be presented or reinstalled in an exhibition context or art related matter. Neither does it address the preservation of a more complex character such as software or Internet based data.

As stated in **Annex 1 – Description of Work in the DCA Grant Agreement:**

“This deliverable will highlight the particular points of interest (for instance the use of open standard formats and codecs) that have to be addressed when creating a strategy for long-time preservation of the digitised content to be implemented in each participating institution. It will also offer state-of-the-art solutions on the use of persistent identifiers to ensure long-term stability of source location for the digital objects.”

There is a distinction between digitised material and born-digital material throughout these *Guidelines*. It appears that institutions that manage cultural heritage make little distinction in practice between digitised and born-digital heritage material, while a large difference exists. Within

digitised heritage collections, the objects still have an analogue equivalent and are distinct entities to which structured metadata can be assigned to. The number of common file formats is limited. Because digitisation can very often be done again, the loss of digitised information is less problematic than the loss of born-digital material. The nature of born-digital material makes long-term preservation very problematic. Born-digital material is produced and distributed in massive quantities, but can also disappear very quickly. It is produced in many formats, made with different software running on different hardware. Many born-digital materials have a dynamic character and do not reach a final status. It's not always clear, therefore, what to archive. What the entity itself is, and the question of what does or does not belong to the object is not always defined either.

Five key digital preservation problems are identified:

1. Abundance of information: how to select what you want to keep? Who is responsible?
2. Object Definition: what is a document¹, or rather an object²? How to define it? What are the significant properties³?
3. Accessibility, logical preservation: how do you ensure that digital information remains accessible in the future?
4. Context: how do you ensure that the meaning of stored information (metadata describing the object) can be understood in the right way?
5. Authenticity and integrity: how do you ensure that authorship and content will not be affected?

When working with artworks the originality and authenticity can sometimes be a dilemma compared to the preservation of any kind of regular text document for example. The reason for this is that there is a risk that the artwork might change during the digitisation and digital preservation process and the artist's intent can get 'lost in translation'. For the artwork, the context and media is just as important as the content. For a regular text document, the text itself is the important part, and not so much the media it was written on. When preserving artworks, this issue is often dealt with by documenting the original artwork, by describing how it looked and what the purpose was, but also the materials used and the technical components needed to display the artwork. It is therefore just as important to preserve the documentation as it is to keep the artwork.

For example, there's a difference between a painting and its digitised version. The digitised version still shows the original image, but it does not show the brush strokes in relief, nor does it give a feeling of the actual size. Another example is a digitally recorded video artwork: the original format worked on a specific platform at the time of creation, but today this platform might not be available anymore. When born-digital media is converted to work on a modern flat panel monitor instead of a bulky cathode ray tube monitor, the images and the story can be seen, but the history and original context is lost. With artworks there is often an extra layer of information that needs to be preserved alongside the actual work. How can one change such an important concept of integrity to an acceptable level, while ensuring that the functioning, the concept, the materiality, the behaviour of the work, as well as the experience and the aesthetic properties are not affected? Some change or loss is inevitable. The question, when digitising media art or preserving born digital art, is: are the significant properties of the artwork preserved, even when others are lost?

The following methods and recommendations all question how to make sure that the digital data can work on future platforms. It is the responsibility of the collecting institutions themselves to take into account how their digital data might be transformed throughout the preservation process, how such changes are related to the understanding of the artwork and how they influence the perception of the work.

1 In the DCA project a document can be an artwork or a contextual document.

2 This regards what a 'digital' object is, contrary to a more bound 'analogue/physical' object.

3 For more information, see <http://www.significantproperties.org.uk/>

The target group for these Guidelines is first of all the participants of the DCA project, but any collecting institution with digital documents, images, video and sound files can benefit from the information.

Ideal solutions are not always possible

Most of the content described in this deliverable *D6.1* tries to give an example of the ideal handling of digital repositories. However, the extent to which an institution can actually follow every aspect depends on its budget, size of repository, and the expertise available. It also calls for sustainable funding, for maintaining equipment, hiring staff with the right knowledge, and updating the preservation plan. Hopefully these *Guidelines* can give an idea of how to get around issues and still have a plan that secures a given institute's repository in the best possible way. NIMk, the leader of WP6 and responsible for this *D6.1*, has for example a good knowledge of the digitisation and preservation of video art with a budget that allows for some equipment, but not for unlimited possibilities. NIMk still has to compromise when it comes to certain aspects of preservation in order to have affordable solutions. For example, NIMk has two preservation copies on one type of hardware instead of the recommended three. NIMk uses proprietary file formats because the best equipment it can afford, supports these formats. The justification for this is that the used file formats and codecs do have some of the other recommended properties, such as notoriety and that it's widespread. This solution still results in high-quality preservation files, although it does not meet the ideal. It is important to have alternatives and to discuss the benefits and shortcomings of every solution in order to obtain a good plan, regardless of limits.

We have outlined what an ideal preservation practice is in this deliverable *D6.1*. Unfortunately, due to different kinds of limitations, it is not always possible to reach this ideal practice because one needs to compromise. The bottom line should be clear: if one is not able to comply with the ideal/best practice, one should be able to explain clearly why one has chosen to do things differently.

3. Definitions of digital preservation

In order to avoid confusion, we will apply the following definitions (which we have created ourselves, unless otherwise stated), throughout the report.

1. **Long-term preservation:** 'Long-term' implicitly means as long as possible, not just twenty years, but hundreds of years. When creating a long-term strategy, it is thought to be a changing and developing procedure, which has to be constantly renewed. As we can't predict the future, a strategy made in 2000 will have to be adjusted along the way and will probably be different in 2012, so when using the expression long-term it does not refer to a permanent solution. Often a timespan of two generations is used to determine if a strategy is long-term. The first generation still has a direct link with the creators of the data (which might help them to access it); the second generation does not. If the necessary measures are taken to make sure that this second generation can still access the data, chances increase that it will survive long-term – especially if each generation approaches the preservation of data in the same way.

2. **Digital content:** The digital content that an institution might have will be referred to as repositories, archives or collections. The digital content regards all digital files an institution uses, administers or distributes, both digitised and born-digital. In these *Guidelines* it stands for everything from audio-visual content to images and text documents. Digital content can also refer to more complex data such as software and net-based art. Since software and net-based art are not part of the scope of the DCA project, we will not be discussing them here.⁴ The digital components will be called digital objects, items, files or data.

3. **Digital Object:** A digital object is defined by the National Library of New Zealand (NLNZ) as being:

- A single file (such as a text document);
- Several connected files (such as a database or a website);
- A collection of independent files that are kept together (such as emails or blog posts).

In these *Guidelines* we will only talk about single and connected files as digital objects, as they are the most relevant.

4. **Preservation method:** a theoretical description of how to preserve digital content in a specific repository. In order to preserve readability and accessibility of the given content, appropriate methods must be chosen, e.g., migration to a specific (new) format every five years, and this in consideration of budget and available technical expertise.

5. **Preservation planning:** a documented policy that should be created for any institute with digital repositories; it includes preservation methods, file formats, and strategy for future actions. The plan should also describe the procedures, equipment, and software needed while keeping data accessible but also authentic in terms of the original intent.

6. **Open standard and open source:** These terms will often be used when addressing file formats and codecs. The difference is that an open standard is an openly described file/container or other where the technical specifications have been described in detail, often by a big organisation such as the ISO (International Organization of Standards). The standard can be freely used, but not altered. An open source format or software is freely available for anyone to build on, use and alter according to their needs. This means that it is available to all in terms of use and openness, but there can be a license fee. Open source is usually not regarded as a standard in itself, but it stands for the idea that the technicalities should be described and open to anyone, without any patents or proprietary properties. Open source software is often distributed with so-called open licenses, such as creative commons or a General Public License (GNU).

⁴ For more info, see <http://www.DOCAM.ca>

4. Trustworthy repositories and file authenticity

As mentioned above, the goal of digital preservation is not only to keep the data accessible but also authentic in terms of original intent.

Gladney illustrates in two of his articles how confusing and difficult it can be to keep digital objects trustworthy and true to what the intent of the producer was, especially long after he/she is gone (Gladney 2005, 2006). This goes both ways. When producing the data, it helps to consider what the end-user would be interested in or would need in order to understand the original. What happens after we have digitised, migrated etc. an object several times? Is it still close enough to its original version?

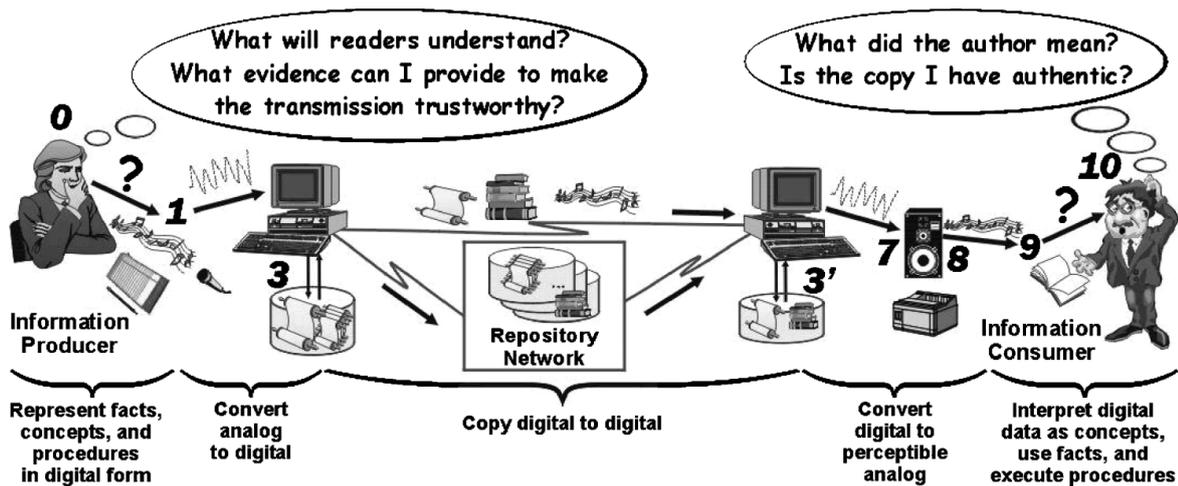


Figure 1. Gladney 2006.

To keep a digital object authentic and trustworthy many of the methods and file formats described in these *Guidelines* need to be implemented. But before that can be done an authentic digital object can be defined, ideally, as the following (Gladney et al 2005):

- Every copy of the original has survived, because of interest in the information;
 - Authorised consumers should be able to find and use any preserved record as its producers intended, doing so without impact from errors introduced by third parties;
 - Necessary information about the object is available, so the consumer can decide whether a preserved object is sufficiently trustworthy for the intended application;
 - Procedures for preservation and copying is automated when possible, to avoid human error.
- The less the human intervention, the more a digital object is likely to be authentic.

The authenticity of files is also very much correlated with the preservation plan. If the right precautions are taken into account when the plan is created, the file will be kept as authentic as possible for a longer time. Information on the choices to be made and the methods to be used follows later on in this deliverable *D6.1*.

5. Standardisation model OAIS

An introduction to digital sustainability often includes the OAIS (Open Archival Information System) reference model developed by NASA in 2002 (CCSDS 2002). The model describes all entities and workflows needed for preserving digital repositories. Digital sustainability is only possible by integrating procedures at every stage, from when a file is ingested into a repository to when it is accessed by a consumer.

The model also gives a description of what a digital object should contain. The digital object itself is called the Content Information (CI) and information needed to preserve the file is called Preservation Description Information (PDI). PDI ensures that the CI is clearly identified. These two parts make up the Information Package (CCSDS 2002 p.2-5). Descriptive Information is then needed to discover and find the packaging information, making the CI searchable. Below there is a short description of the different entities in the OAIS model. They will be described shortly in the following, with examples on how this can be interpreted for use in the DCA project.

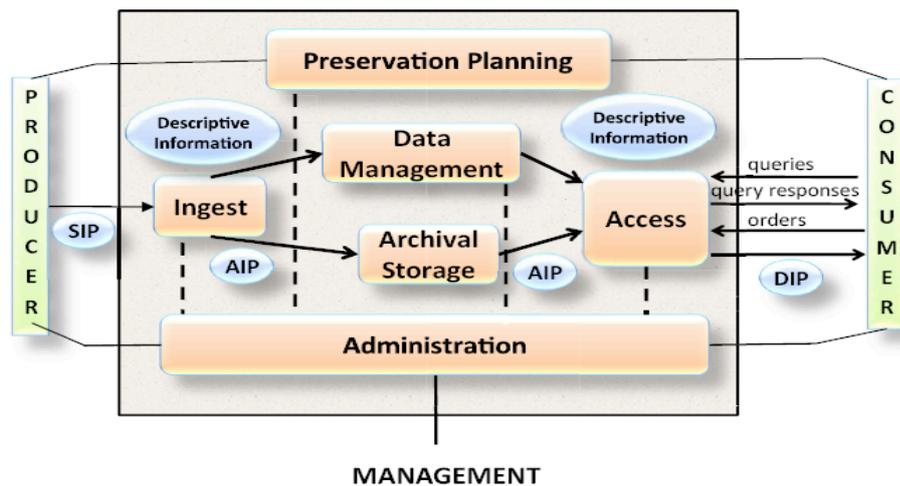


Figure 2. OAIS Functional entities (CCSDS 650.0. 2002)

Ingest is when the digital object is received from a producer as a Submission Information Package (SIP). These packages are prepared for storage by first ensuring that the institution's demands in terms of what type of files they receive are upheld. This is partially achieved by quality control, where the digital object is checked for errors and the content is as expected. Then the package of information is prepared for archiving by being made into a so-called Archival Information Package (AIP) consisting of the information from the SIP and the Preservation Description Information (PDI) that describes how best to preserve it.

DCA example:

After digitisation the digital object should be validated for quality assurance, in order to see if the file is error free and has the expected properties. Specific tools for file validation and metadata needs to be used. Basically the file should be checked to find out whether it can be opened and that file naming and other information are the same as before digitisation. This is not usually done manually (one file at a time), since it would be time consuming and entail a risk of human error. Extraction of metadata and tools for this are explained more thoroughly in *Chapter 11. Preservation metadata*. On the subject of preservation metadata, before ingesting the digital object (Information Package) into archival storage, one should register preservation metadata that are suitable for each object. This is basically information

needed in order to read and access the object in the future. By making a database with these metadata connected to the actual digital file, the information can be safeguarded so that if the file gets lost or corrupted, it can explain how and what it actually contained (See more in *Chapter 11. Preservation metadata*). One should also create a unique, persistent identifier (see more in *Chapter 13. Persistent Identifiers*) and create a checksum for every file (see *Chapter 12. Checksums*). This is to safeguard the authenticity of each file and to avoid changes going unforeseen. This can be done manually or automatically. The checksums could be made part of the file's unique identifier. When these precautions are made the digitised object is ready for archiving.

Archival storage is the entity that receives the AIPs from ingest and maintains them for permanent storage. This means preservation of the raw bit data, which includes maintaining or refreshing hardware, error checking and disaster planning (backup plans). This is important when preserving raw bits, because without them there could be no access to the original content. The more practical aspects of this process will be described thoroughly in a following section on establishing a digital archive.

DCA example:

When all preparations for each object have been made during scanning and digitisation, they are ready to be stored in the archive. This is where the data is physically stored on hardware. The recommended hardware for long-term storage are hard disk drives (HDD) and archival storage tapes, better known as LTO (Linear Tape Open) (Bradley et al 2009). The reason for this is that these storage types are relatively stable and can contain a lot of data. They also present the possibility of setting up large server systems to retrieve data readily, which makes it easier to maintain and check for errors automatically. If one chooses an external cloud storage host, it will most likely also be using hard disks as servers. Do keep in mind that although these storage devices are recommended, they still have to be replaced after between 5-10 years according to the vendors, to avoid any technical or mechanical failure. Usually new versions of hardware with more storage space appear every other year. It is recommended to check the data on the hardware every six months, to be sure that the bits are not changing at all and that files are still compatible with the available equipment.

It is not recommended to use CDs, DVDs or Blu-Ray discs for long-term storage, because the stability and risk of error or damage is greater, during the making of the disks and during handling and burning. Also the amount of storage on an optical disc is small and it can be difficult to retrieve and access a growing collection in a useful way. If this medium is used for preservation purposes, make sure to have several copies and follow the advice given by JISC (www.jiscdigitalmedia.ac.uk/stillimages/advice/using-optical-media-for-digital-preservation/). Solid-state drives (SSD)⁵ are also not recommended because their stability has not yet been fully researched and at the moment they are expensive. One test, which was conducted on three different types of SSDs, showed that although an SSD is more effective than an HDD on some levels, there are still some elements that need improvement before it can be recommended for preservation purposes (Chen et al. 2009).

Archival storage is also about backup plans and disaster planning. This means that for each information package stored we need to have at least three copies of the same data. These copies should be on at least two different types of hardware and in principal also located at several geographically different places. Usually two to three locations are recommended. Of course this

⁵ A solid-state drive (SSD), sometimes called a solid-state disk or electronic disk, is a data storage device that uses solid-state memory to store persistent data with the intention of providing access in the same manner as a traditional block i/o hard disk drive. SSDs are distinguished from traditional magnetic disks such as hard disk drives (HDDs) or floppy disk, which are electromechanical devices containing spinning disks and movable read/write heads. In contrast, SSDs use microchips that retain data in non-volatile memory chips and contain no moving parts. (http://en.wikipedia.org/wiki/Solid-state_drive)

doesn't mean that there isn't still a risk of damage to the data, but at least the risk is minimised. External providers can supply such precautions. There exist many providers, but it's important to know what they actually do to safeguard your digital collections. You will find more about this subject in *Chapter 7. Bit preservation and logical preservation*.

Data Management stands, in short, for updating and administering the databases of both the descriptive information and the information packages themselves (mostly AIPs). This entity is closely linked to Archival Storage, because it keeps track of all the AIPs in the storage system.

DCA example

The data that is stored for long-term archival storage has to be searchable and retrievable, in order to keep track. This requires a database that is outside the archival storage, and which contains descriptive information linked to each and every object in the archive. This information is called metadata and whenever there is a new object, details about it will be added to the database. Also when transforming the objects, for example migrations or other preservation procedures, there needs to be a registration in the database of such changes. Staff in connection or within your collecting institution will be needed to build and maintain the database. To read more about preservation metadata and which systems can be used to create a database, see *Chapter 11. Preservation Metadata*.

Access is the entity that receives requests from end-users and provides access to the information stored in the archive. The AIPs from the archival storage are made into Dissemination Information Packages (DIP) before the information is delivered to the end-users who have requested the package. The DIP is made to meet the customer's requirements and while controlling and limiting access to specially protected information (CCSDS 2002).

DCA example

The reason for safeguarding data for the future is that someone might, hopefully, be interested in the information. The end-users and consumers in the DCA project will typically be the public interested in the given institution's art collection. The project's goal is to make such collections available online for a broader public. But in the process it is important to focus not only on the creation of data from the collection that is suitable for the Internet. If the previous points are followed, you will have an information package containing information suitable for long-term storage. A suitable copy can then be made for each of these storage versions (for suitable preservation formats and codecs see *Chapter 8. Introduction to recommended file formats and containers*). Whenever a file is requested or needed, it should be accessible through archival storage and copied into a version suitable for the Web or publication, depending on the request. In the DCA project all collecting institutions have selected the works that they want to make available through Europeana. They can therefore make Web and publication copies at the same time as the preservation version, without having to retrieve the preservation version from archival storage first.

Preservation planning concerns preservation of accessibility and readability of data. The functions of preservation planning address recommendations for file format standards, monitoring changes in technology, evaluating content of a digital archive and making an overall policy on the subject. The preservation plan should reflect what current strategies would preserve the access of content in the best possible way. Selecting a suitable solution, by using different tools, makes it possible to implement a specific method (for more information on these methods, see *Chapter 9. Methods for preservation*). Another important role for the preservation planning entity is the definition and monitoring of the designated community. The designated community is a profile of the end-users of digital objects and of the type of information they typically request and know about. For example, they might want MP3 files instead of WAVE files, or a whole new file format might become increasingly popular. Whenever there is a change in their demands, the preservation planning entity should consider a strategy update to meet such requirements without

compromising the security of the original content. Below, a tool recommended for making a preservation plan will be introduced (*Chapter 10. Preservation planning*).

Administration should provide the overall policy for the archive. It defines which standards to accept and what licenses and rights the archive has. It also has the responsibility of actually maintaining the hardware and software, as well as updating archive contents (CCSDS 2000). This entity is correlated with preservation planning, which sends suggestions on how to preserve access to digital content. The administration can then make their policy and recommendations based on such information.

Overall

All of the above-mentioned entities are more or less correlated and if one component doesn't work, it might put the preservation or future readability of the content at risk. The OAIIS, as a model, addresses what an archive should be able to distribute and maintain. It also gives a definition of what digital preservation planning actually is, but it does not give specific guidelines on how to set up a workflow or what formats or methods are best. The interpretation and implementation of these components can therefore be done in many different ways. The OAIIS model shows the ideal situation and is therefore not always fully applicable for all institutions. In the following chapter we will attempt to describe which preservation strategies and solutions are recommended.

6. How to get started

First of all, it's important to have someone within the organisation who makes sure that a preservation plan is created and followed through (RLG 2007). The person responsible should also make sure that digital preservation is maintained, as it is an ongoing and permanently developing process. Before making a preservation plan, it is important to assess the repository and the needs of the organisation. In order to implement a long-term strategy, it must not only address what specific method is used, but also describe when and how to renew and re-evaluate the chosen method and the repositories, in order to diminish the risk of content or hardware becoming obsolete.

Some of the first questions could be (based on Becker et al. 2009):

- Who is responsible for the preservation and the maintenance of the digital repository?
- How large is the repository (GB – TB)? → how much storage space is currently needed and how much storage space will be so in the future?
- Which sort of files does the repository contain? → Are they useful for preservation purposes or should they be changed?
- What are the available resources? → Can we buy the best equipment or do we have to settle for the next best thing?
- Is there financial support for maintaining the preservation strategy? → Can we get funds for more storage and maintenance when we have digitised the collection?
- Who is going to use the files? → Is the information for internal, public or specific research use? Do the end-consumers have specific demands regarding the quality or format?
- How are the files going to be accessed? → Through the Internet, or only on command? How can the connection between archival storage and accessible files be made?
- Who are the producers, and what demands and responsibilities can be placed on their shoulders?

All of these questions are relevant, especially when starting a digitisation project. If objects are being digitised, the resulting digital copy should reflect the best possible version of the original, not just for the first five years but also for many years to come. It is therefore very appropriate to start making a preservation plan even before the real digitisation starts.

To obtain an overview of the areas in which an institution needs to acquire more control, the Digital Preservation Capability Maturity Model (Dollár et al. 2009), can be used as an easy, basic start. It resembles the short survey that all participants did at the beginning of the DCA project, but goes a little further. There are fifteen topics on the basic levels of preservation, ranging from strategies, technical expertise, storage management, digital preservation metadata, and accessibility. By filling out the model (giving between one - four points on each topic) an institution can obtain an estimate of how advanced, and how capable the organisation is for making a preservation plan. The model ends with a numerical result classing an organisation into one of four levels from 'nominal' to 'advanced'. To fill in the Digital Preservation Capability Maturity Model, check *Appendix 1*.

The Digital Preservation Capability Maturity Model in itself does not create a preservation plan, but it points out the areas a given institution should start to improve. For example, a small institution lacking the technical expertise should consider either hiring someone internally or contracting an external company or institution.

6.1 Topics of the Digital Preservation Capability Model

1. *Designated Communities.*

Concerns the institution's users and their needs, as well as how the collecting institution receives and accesses files.

2. *Collaborative engagement*

Relates how the collecting institutes collaborate with other companies. Whether, e.g., they use existing models and standards or do new research in order to overcome common preservation problems.

3. *Governance Through Identified Roles and Responsibilities*

Questions whether there is an agreement within the programme on who is responsible for the preservation of the digital repositories, and which roles there are to fill within the institution. It also looks into whether such responsibilities and roles are written down.

4. *Policy*

Asks whether the collecting institutions have a written policy for a preservation plan, and what the goals and the need for it might be.

5. *Strategy*

Looks into whether there is a written strategy describing actions and procedure: how to keep the bit stream "alive" and accessible with planned backup and renewal of media. Does the strategy also include the implementation of any long-term plans, such as migration or emulators made for specific software and digital repositories?

6. *Digital record survey*

Questions which files the collecting institution needs to preserve and whether they have an overview of how to locate them. The files can be categorised into three different types:

- Legacy digital records: Digital records embedded in obsolete formats that can only be extracted by special computer de-coding and re-encoding.
- Near-legacy digital records: Digital records embedded in technology dependent formats, but still possible to migrate.
- Non-legacy digital records: Digital records embedded in open source, technology neutral formats with backwards compatibility.

7. *Storage Management*

Questions whether there are multiple copies of the data, whether the copies are stored on different equipment, and on different geographical locations. How are these storage facilities maintained, interlinked, and so on?

8. *Digital Record Ingest*

Investigates whether the institution follows OAIS recommendations, according to ingest. Does the institution have virus checks? Does it have file validation and normalisation? Does it use metadata capturing and embedding for all new or incoming files? Does the institution have metadata available that at least provides context regarding the creator who uses the digital records and any possible relationships with other digital records?

9. *Digital Record Security*

Relates to the collecting institution's processes of:

- Technically blocking unauthorised access to digital records;
- Periodic backup of digital records that are stored at off-site storage repositories;
- Disaster response and business recovery.

10. *Planned Device & Media Renewal*

Looks into plans for renewing hardware on a regular basis and who is responsible for maintaining the equipment.

11. *Technical Expertise*

Questions whether the institution has enough technical expertise to make a preservation plan. If not, how can that knowledge be obtained? Should it come from external or internal staff?

12. *Access to Digital Records*

Investigates how files that are stored for long-term accessibility, can be accessed.

13. *Digital Preservation Metadata*

The preservation programme should include procedure for metadata. This also addresses whether the institutions have a metadata database, how it is collected and what it includes.

14. *Digital Record Integrity*

This addresses whether there is a procedure for regularly checking file integrity, especially after migration or conversion, or other file transformation, i.e., comparison of encryption (like md5 hash and sha-2).

15. *Open Source and Technology Neutral Open Standard Formats*

Questions whether the used software and technology is open source, and whether it supports open standard and technology neutral formats while incorporating extra software that can monitor the obsolescence status of file formats.

Evaluation of results

After making a survey of the institution's repository and the present level of use and capability, it is time to consider what exactly to do. The different approaches will be described in the following sections. If the collection institution is on a 'nominal' level, there is good reason for it to start from scratch: to make a preservation plan, evaluate the repositories' size and type of content and to figure out if it can hire staff to maintain the collection or if it needs external advice. It may be a good idea to make the most of this guideline. If the collection institution is in one of the 3 other levels, the model gives an idea of which areas are not up to standard. Once one has completed the model and can see which questions one has answered, the table below (Table 1) can be used to see which areas of the guideline will be helpful.

Questions with a low score	Go to chapter:
1 – 4	5. Standardisation OAI model, 9. Methods for preservation, 10. Preservation planning.
5,8, 9,12	7. Bit preservation and logical preservation 9. Methods for preservation.
6	7. Bit preservation and logical preservation, 8. Introduction to recommended file formats and containers.
7,10,14	5. Standardisation OAI model, 7. Bit preservation and logical preservation.
8	5. Standardisation OAI model, 11. Preservation Metadata.
13	11. Preservation Metadata.
15	8. Introduction to recommended file formats and containers.

Table 1

7. Bit preservation and logical preservation

There are several approaches to a digital preservation strategy, but two overall aspects must always be taken into consideration: bit preservation and logical preservation. Both are explained below.

7.1 Bit preservation⁶

An important part of long-term accessibility takes into account the possibility of equipment failure, accidents happening to the main server or maybe even a building (like for example a fire). A minimum requirement is that important files should still be available somewhere else. These files should not be edited or transformed; otherwise the authenticity is at stake. They should be preserved just as they are for later access and use.

In order to maintain the file bit sequence, a preservation strategy and plan must include bit preservation. As the name implies, it is important to make sure that the bits remain intact and in the right sequence. It is recommended to use at least two different types of hardware for storage. If one hardware type becomes obsolete or the equipment breaks down, there should be another way to access the content. It is important that backups are stored in geographically different locations in order to be able to retrieve the information from elsewhere. A system that maintains and checks the bit stream⁷ is needed in order to avoid bit rot⁸ or unexpected changes. This should be a regular and automatic procedure throughout the year. The hardware itself should be regularly checked and exchanged for new versions every couple of years, depending on what is recommended by the hardware vendor of that specific equipment. The ideal system is one that alerts one if a hardware component is not functioning correctly, and allows the failing component to be replaced immediately. Keep in mind that even this kind of system does not solve the issue of obsolete equipment.

Checklist for bit preservation:

- Multiple copies of each digital object (at least two, the more the better).
- Kept on multiple media storage types (at least two different types).
- Different copies stored in geographically different locations (three is recommended).

Backup procedures immediately after an object is transformed.

- Hardware and software maintenance after each new plan, re-evaluation and update of that plan every five to ten years.

It is crucial to implement quality control and regular monitoring in order to detect any errors happening, either due to the hardware itself or when moving or changing files. This control and monitor procedure should be done on a regular basis (for example every six months) to avoid problems at the moment when the files are needed. It should also be done during digitisation projects to ensure that the resulting digital files meet the quality expectations.

7.2 Logical preservation

The process of preserving file accessibility and ensuring that they are still understandable and readable, regardless of the evolving technologies, is referred to as logical preservation. In order to do it properly, there are some basic procedures that must be taken into account:

- A choice of formats and codecs wanted in the repository, and an overview of which software supports them;

⁶ Bits = a unit of information expressed as either a 0 or 1 in binary notation. Bit preservation is maintaining the order and character of bits of which a given file consists.

⁷ A stream of data in binary form, basically any file is a bit stream.

⁸ When one or more bits in a given file changes, i.e. a zero becomes one. This can result in the data not working as intended.

- The production of metadata for every file: depending on the size of repository and type of data the metadata can be kept in different ways. With a big collection of files it is recommended that one stores metadata in the ideal situation i.e., basic information embedded within the file and more complex metadata in a specific database linked to the files. With smaller collections it is sometimes better to keep the metadata for each file embedded within the file itself (Bradley et al 2009). ;
- Regular evaluation of the software and file formats and codecs to avoid obsolescence. This can involve a broad range of approaches. Each approach varies depending on the repositories' content.

Digital preservation can be approached in many ways by using a number of different tools and strategies. *Chapter 9. Methods for preservation* will address descriptions and recommendations for different ways of preserving the access to digital content. If the data is in some way obsolete, it becomes increasingly difficult to process the information. It is therefore necessary to monitor and check data and implement preservation strategies before it gets too time-consuming and difficult. This implies that bit preservation alone will never be sufficient in the long run for the preservation of data because technology will always inevitably change. A proper preservation plan will only be complete if it includes both bit and logical preservation.

8. Introduction to recommended file formats and containers

Before the start of any digitisation project, the choice of file format should be carefully considered together with the available technology, storage facilities and in-house versus external expertise. The type of formats for digital preservation will be elaborated upon in these *Guidelines*.

8.1 Codecs and Containers

One often comes across the terms 'containers' and 'codecs', especially when working with audio and video formats. In order to understand such terms in this context, a definition is provided below.

Container: In order to view a video one needs a video and an audio stream. These two streams need to be packed together within a container. Depending on the format of the container the information can be packed differently. It can be layered, meaning that within the container the audio stream is separate from the video, or the audio and video come together in one mixed stream. A container can also be referred to as a wrapper. A file format can be a container format, i.e., it can contain different audio and video streams.

Codec: is short for coder/decoder, meaning it can be any technology for coding and decoding data. For audio-visual material this type of coding/decoding often includes some form of compression and decompression. If a video is not compressed, it requires a huge amount of storage space and so to avoid a huge bandwidth during streaming the material needs to be decompressed when played. Depending on the codec, the compression/decompression is done in different ways. A video or audio stream is for some reason often referred to as a codec, because the specific type of stream requires a specific codec to be played, but it is not in itself a codec. In these *Guidelines* a codec refers to the programme or other technology that codes/decodes a given video or audio stream and not to the stream itself.

A **file format** gives one information about what the file is (audio, text etc.) and how the data is stored (uncompressed, lossy etc.). A file format can be utilised coded/decoded by its own, specific codec made for that type of file format, or if the file format is a container it can contain different audio and video streams that require different codecs. Some video and audio versions cannot be used alone, they need to be within a container in order to be streamed.

For example NIMk uses the AVI (Audio Video Interleave) multimedia container. The container can contain both video and audio data. Within the AVI files used by NIMk there is a V210 video stream coded with a v210 and a Pulse Code Modulation (PCM) audio stream that also needs a specific PCM codec. A file format and a stream can be labelled as either uncompressed or compressed. The differences will be explained below.

8.2 Uncompressed, lossless and lossy compression

Uncompressed files are basically the original information that was available when the file was made. When recording audio, video and still images digitally, one can either choose a format that does not compress any information or one that does. Having an uncompressed file means that all the data in the video or still images are intact. This means that the file contains the richest data possible. This richness offers a broad range of possibilities for manipulating the information, like adjusting colour balance and so forth during the editing process. For example, when taking uncompressed photographs with a professional Canon camera the resulting files could be stored in the RAW format. This is a file format that cannot be stored as a viewable file; it can only be used for editing. To preserve the information completely intact, it is recommended to convert this RAW file into an uncompressed TIFF file. This means that the data is not compressed when the file is stored (and there is no loss of data on colour, brightness etc., due to certain forms of compression). Thanks to this the TIFF file still holds a broad range of editing possibilities. If the

image is saved in a lossy compressed format such as the JPG format, there are less editing possibilities because the file can no longer hold the same amount of information anymore.

Compression is primarily used to reduce the file size and transmission speed. Audio, video and still images can be quite large in file size, and therefore slow to work with. A large file size can for example result in very slow loading. As such it can be useful to compress a file. There are different types of algorithms used for this process. In general one can make a distinction between two versions: lossless and lossy compression.

A **lossy** format is a compressed file that has some loss of original information, due to attempts to minimize the size of the file. The result of this loss of original information can sometimes result in a bad/pixelated image quality in photos and video, or in a bad sound quality in audio files. When digitising, we have the choice of making an uncompressed or compressed file. It is not recommended to create lossy compressed files because this can result in a loss of information and thus a lower quality. The question to be answered is: is any essential information lost during compression? We don't know how technology will be evolving. Maybe we can't see today that important information is being lost during compression, but this loss will become apparent when new technology is available. The format should preferably be uncompressed (or lossless compressed) to ensure that as much information is preserved as possible. An exception might be the case when only lossy compressed originals are available. If one receives an original digital object that is already in a lossy compression form, it should not be changed, but instead kept in that lossy compressed state. Converting it to an uncompressed version won't add any extra information and the whole procedure will be futile. Other examples on when to compromise with uncompressed files can be seen at the end of this section.

A **lossless** format will compress the file information, but it maintains the possibility of retrieving original information through the use of algorithms. By choosing the right lossless format, the storage capacity can be utilised more efficiently than with uncompressed files, without visibly reducing the quality of the files. In most cases uncompressed is recommended above lossless compression, because every time the file is compressed and decompressed there is a risk of something going wrong. Therefore, unless the content requires large amounts of storage, for example audio-visual material, it is recommended that one continues using uncompressed formats. Some have mentioned the use of a near-lossless format. This kind of format would offer the possibility of reconstructing a lossy file to an uncompressed version. But whether this really works is still under discussion; a file might lose some of the information and not be perfectly restored to its original condition (Schmalen et al 2009).

Which type of file a given institution should use depends on several things. If the data file is strictly stored for preservation or for publishing and printing purposes, it is recommended that one stores the file uncompressed. The reason for this is that an uncompressed file offers the best possible quality and a suitable copy can be made from it. If the institution has a small storage capacity or has very large amounts of information, it might become impossible to opt for an uncompressed file format and necessary to store the files in a lossless format. This was the case for the British Library that digitised its whole newspaper collection and stored the files in JPEG2000 format. There were two main reasons for the British Library to consider it unnecessary to store the digitised versions in an uncompressed file format (DPE 2010). First of all the number of newspapers to be digitised was very large. Furthermore, the goal of the digitisation process was to make the newspapers available and searchable in an online database.

One could also imagine the use of a lossy compressed format when the files are destined for web use only. File size can be reduced for quicker and smoother access, because often the content doesn't have to be in the highest possible display resolution, but the capacity on the Internet also changes. For example, a few years ago videos on the Internet would often be of a low quality and small size, but today a lot of videos are often streamed in HD. As time goes on, the files are

getting bigger and of a higher quality while the streaming is getting better and faster. Having a good source is vital.

The ideal solution could be to preserve two versions of the same file: one lossless user copy and one uncompressed preservation copy (Brown et al. 2008). Unfortunately, this is not always possible due to limited storage space, especially for video material. Other reasons not to choose the ideal, uncompressed formats are:

1. If the source is already a lossy compressed file. For example if an original artwork image is a JPG, there is no need to save it as a TIFF because it will not improve quality. Information that is not yet part of the source file cannot be added so to speak.
2. If the source is low-quality. For example when digitising a VHS tape, the quality of the source is already low. One cannot gain any more details by making it into a high-quality uncompressed digital file.
3. If one doesn't have the facilities. For example when one digitises videotapes but can't afford the highest quality equipment, one might have to compromise and take the "next best thing".

8.3 Recommended file format properties

Choosing a suitable file format for digital preservation is different in terms of quality than preparing a file for presentation or web applications. For digital preservation one should choose the best suitable format to derive file formats for different kinds of presentations. You can derive lower quality, seldom higher quality.

When to choose a format or codec suitable for long-term preservation:

1. When starting a digitisation project;
2. During ingest: when a collecting institution receives new files, the files can be normalised (as will be explained in the section about methods for preservation);
3. When the collecting institution produces its own born-digital files.

A file format should have some specific properties regardless of what type of content the format holds. Recommended properties for file formats based on Brown et al (2008) are listed in table 2.

Ubiquitous	The file format should be widespread, often used and well known.
Open standards	An acclaimed standard with available technical specifications.
No patents	Formats without patented technology or licenses are preferred.
Metadata	The possibility of embedding unique identifiers and other metadata within the file.
Multiple view-paths/Support	There should be more than one type of software for visualising or rendering the file format.
Uncompressed formats	For preservation of all data, the use of uncompressed file formats is preferred, in order to keep the best possible quality. Depending on loss, use and quantity, lossless compressed form can be an option. ⁹
Stable	No major or constant changes and also possible backwards compatibility with older versions.

Table 2

⁹ For instance if one has the choice to reduce one's storage needs by 90% (as is the case with JPEG 1;10 compression) or more with little or no discernable loss.

One should be aware that when a format has one or two of these properties, it is not necessarily a good preservation format. However, the more the above properties a file format has, the better it is for long-term preservation. For example, the MPEG video and audio formats are very popular and often used in all kinds of institutions. They are acclaimed open standards and a lot of software supports them. However, this widespread use does not make them suitable for long-term preservation. MPEG may be an open standard, but it often requires a license or makes use of lossy compression. *Section 8.3* gives a selective overview of recommended formats for long-term preservation of different types of content.

8.4 Recommended file format standards

The following tables show a description of file formats recommended for digital long-term preservation. The formats have different properties and functions, but all of them fulfil most of the recommended properties. The American Library of Congress has made an online digital format dictionary that contains short descriptions of each format as well as summaries of sustainability factors (Library of Congress 2011). Cultureel Erfgoed Standaarden Toolbox (CEST www.projectcest.be) and Digital Heritage Netherlands (Digitaal Erfgoed Nederland, DEN) also have useful overviews of standard formats categorised in audio, video and so on. They also provide links to relevant literature on formats. This is very handy when considering the properties of a file format. These sites have inspired the lists included below (Table 3-6). One should keep in mind the fact that not all lists on the websites take into consideration which formats are good, specifically for long-term preservation. The Library of Congress sometimes only states the facts and technicalities of a given file format, and DEN describes the ones that have been acclaimed standards, but not specifically for preservation purposes. In the following tables one can find a list composed of file formats that are specifically chosen because of their good properties for digital preservation. File formats that are better for presentation or web purposes can be found in appendix 2.

The last column of the tables is called “sustainable qualities” and refers to the file formats’ properties that correlate with good digital preservation practices and qualify them for long-term storage. File formats can either be open standard or open source, as described in the introduction. If it is a standard format, it is well defined and the technicalities are available but can’t be changed. If it is an open source format, it is made in an open community and can be altered or changed but one needs to take into account that some might have certain open licensing principles.

	Text documents	
File format abbreviations	Description	Sustainable qualities
PDF/A ¹⁰	The portable document format (PDF) is a well-known and well-used document format. Adobe relinquished control of PDF 1.7 and any following versions to ISO, who confirmed it in 2008 as an international standard (ISO 32000-1). Since then the PDF was developed with the possibility of embedding metadata within the file, hence the current PDF/A format which is recommended for digital preservation.	Open standard, widespread, metadata embedding, and platform independent.
SGML ¹¹	Standard Generalized Markup Language is a platform independent format for organizing text documents. SGML is the formal basis of HTML and XML.	Open standard, platform independent.
ASCII ¹²	American Standard Code for Information Interchange is a basic	Open standard,

10 ISO/IEC19005-1 (2005). Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF 1.4 (PDF/A-1).

11 Digitaal Erfgoed Nederland, www.den.nl/standaard/6/Standard-Generalized-Markup-Language

	text language that cannot contain graphs and pictures, but encodes how plain text is presented. It is used in file formats such as TXT and RDF. It was made into a standard by ISO in 1975. Compatible with web text coding format UTF-8.	widespread, platform independent.
ODF ¹³	Open Document Format for Office Applications is an open, XML-based format for office files, such as word-processed documents or spreadsheets. It was made an international standard in 2006 (ISO/IEC 26300) and offers a suitable format for the preservation of digital documents created in proprietary office formats like those generated by Microsoft Office (Weil, R. 2011). The file format makes transformations to other formats simple by leveraging and reusing existing standards wherever possible.	Open standard, platform independent.

Table 3

When scanning documents one should have two connected file types, the image master file and a structured text file. The image master is basically the plain image of the document, readable but not searchable. The structured text file is one that has been made searchable and machine-readable, for example with OCR (Optical Character Recognition). This also makes it easier to find a file or search for specific words in a file. The two versions of the document should be linked and preserved, for example with the help of metadata (see chapter 10).

	Still images, photographs, scanned documents	
File format abbreviations	Description	Sustainable qualities
TIFF ¹⁴	Widespread image format made by Adobe, it has not changed since 1992 (only a few adjustments). A TIFF file can be lossless compressed or even uncompressed. It is the preferred image format for high quality photographs because, unlike the JPG format, it can be edited without loss of quality. The TIFF structure allows for metadata embedding of information about the image in the header. TIFF can also have more than one layer for editing in, for example, Adobe Photoshop or it can be used as a container for compressed lossy or lossless image formats. IBM baseline TIFF is recommended for preservation purposes ¹⁵	Open standard, widespread, possibility for uncompressed content, metadata embedding, and platform independent.
PNG ¹⁶	Portable Network Graphics is an extensible file format for the lossless, portable, well-compressed storage of raster images. PNG is designed to work well in online viewing applications, so it is fully streamable with a progressive display option. It is an improved version of GIF, with the possibility of metadata embedding.	Open standard, widespread, lossless content, metadata embedding,

12 ISO (1975). "The set of control characters for ISO 646". *Internet Assigned Numbers Authority Registry*. ASCII is strictly spoken not a file format but a character encoding scheme (like Unicode).

13 ISO/IEC 26300:2006. Information technology, Open Document Format for Office Applications (Open Document) v1.0

14 Adobe (2008). TIFF, Adobe Systems Inc., <http://partners.adobe.com/public/developer/tiff/index.html>.

15 TIFF should, for preservation purposes, only be baseline TIFF and not extended TIFF, because not every TIFF reader is obliged to be able to read all TIFF extensions. RGB images are part of baseline TIFF, but CMYK images are only part of TIFF extensions. IBM indicates that the TIFF is configured for little-endian, which is recommended because the CPU's of most current computers are configured little-endian.

16 Adler et al. (2003). ISO/IEC 15948:2003 (E), edited by David Duce: Oxford Brookes university, W3C.

	Still images, photographs, scanned documents	
		and platform independent.
JPEG2000 ¹⁷	In 2000 the Joint Photographic Expert Group launched a new version of the well-known JPEG. The open source format is said to have a better and different compression algorithm, resulting in a lossless and no longer only lossy compression - although lossy compression is also still possible. Also the JPEG2000 (JPG2) can contain metadata. The format is not as well known and used as JPEG. It ensures that information about the file is still possible to retrieve, but it is not entirely risk free.	Open standard, lossless content, metadata embedding, and platform independent.

Table 4

The reason that DNG (Digital Negative) and RAW are not mentioned in this list is that for the time being these formats are not recommended for long-term preservation purposes. An image saved as a DNG or RAW file contains minimally processed data from the camera sensor, making red, blue and green pixel information. Raw files are so named because they are not yet processed. Raw image files are sometimes called digital negatives, as they fulfil the same role as negatives in film photography. An analogue negative is directly usable as an image, but has all of the information needed to create an image. RAW and DNG are developed for editing purposes and are not compatible with regular image viewers because the information has to be interpolated first. In order to use them, an editing programme such as Lightroom or Photoshop is needed. Also RAW is usually a proprietary format with different properties depending on which camera brand one uses. For preservation purposes a DNG and RAW file would have to be edited or processed and then converted into a more suitable viewing format, such as uncompressed TIFF. This way, the image can still be edited, but it can also be stored for preservation purposes.

	Audio	
File format abbreviations	Description	Sustainable qualities
FLAC ¹⁸	Free Lossless Audio Codec is an open source, lossless audio format. Compared to mp3 files that can be compressed up to 80% of the original, a FLAC file will be between 30-50% compressed (Bastijns et al 2009). All operating systems subsidize FLAC.	Open source, no patents, lossless content, platform independent.
WAV ¹⁹	Waveform Audio File Format is an IBM and Microsoft audio file format that can contain uncompressed audio. WAV uses pulse code modulation, and is a standard according to the European Broadcasting Union. WAV files can be read on different operating systems. Usually WAV is preferred over FLAC, amongst others, because it is more widespread and supports more audio channels.	Widespread, open standard, possibility for uncompressed content, uses PCM, container.
BWF	Broadcast Wave File is an extended version of WAV that can embed metadata in the header and also contain uncompressed audio. It is also developed by the European	Open standard, embedded metadata, possibility

17 ISO/IEC 15444-12 (2005). Information Technology – JPEG 2000 image coding system – Part 12: ISO base media file format. 94p.

18 Coalson, J. (2008). FLAC: Free Lossless Audio Codec. <http://flac.sourceforge.net/documentation.html>

19 An alternative for WAV might be AIFF (Audio Interchange File Format) that was developed by Apple. Like WAV, AIFF uses pulse code modulation.

	Audio	
	Broadcasting Union, and now considered a standard by the Audio Engineering Society (AES) and recommended for digital preservation purposes (Bradley 2009 and Emmett 2000).	for uncompressed content, uses PCM, and container.

Table 5

PCM (Pulse Code Modulation) is a method to digitally represent sampled analogue signals; the higher the sample rate the better the digital sound represents the original analogue signal. The PCM audio stream is recommended for audio and video, but is not included in the list because it does not function alone. A PCM file is the raw audio information (the equivalent of RAW image files) and needs to be in a container like WAV or AVI. If not, it can be difficult to find software that can interpret the data.

	Video	
File format abbreviations	Description	Sustainable qualities
MXF ²⁰	Material eXchange Format (MXF) is a container format for professional digital video and audio media, defined as standard by the Society of Motion Picture and Television Engineers. It can contain metadata in the header and is made for keeping video and sound files together.	Open standard, possibility for uncompressed data, embedded metadata, container.
AVI ²¹	Audio-Video Interleaved is a container format developed by Microsoft. AVI containers can hold audio and video content with different bit-rates and frame-rates. The header can contain information about the video, such as frame-rate and width (metadata) (Bastijns et al 2009).	Widespread, possibility for uncompressed data, embedded metadata, container.
MOV	Multimedia format, made by Apple, originally for QuickTime frameworks. It is a proprietary container, but is widely used for video and audio content.	Widespread, possibility for uncompressed data, container.
Mjpeg2000 MJPEG2000 ²²	This video stream and container format consists of JPEG2000 images for every frame. In order to have a video with audio as well, the MJPEG2000 must be combined with an audio format within a container (for example MXF). It can also be used as a wrapper.	Open standard, lossless images, platform independent, container.

Table 6

There is no widespread consensus on a preferred preservation format for audio-visual material. Therefore, it can seem like the list is incomplete or not consistent. This overview list some video formats and containers that are neither standards nor open source but they offer uncompressed video and their use is well spread and well supported. At the same time several types, such as

20 P. Ferreira (2010). MXF – A technical overview. EBU Technical review (Online). P .
http://tech.ebu.ch/docs/techreview/trev_2010-Q3_MXF-2.pdf

21 Bastijns, P., Coppens, S., Corneillie, S., Hochstenbach, P., Mannens, E., van Melle, A. (2009). BOM – Vlaanderen (Meta)datastandaarden voor digitale archieven. Universiteitsbibliotheek Gent. P.71-177.

22 ISO/IEC 15444-3 (2007). Additional profiles for archiving applications.

MPG2, are not mentioned in this list, but they are often used for digital preservation purposes because they are well known. The reason they are not on the list is related to proprietary issues, such as license agreements. MPEG2 cannot be used for uncompressed or lossless compressed files. Then there is also an open source lossless video codec called Huffiyuv, which might be a better option. But at present Huffiyuv is not very widespread. The discussion could go on, but ideally the file formats and containers chosen should have all recommended properties stated at the beginning of this section. For audio-visual material, this means that the choice should be a MXF container with a MJPG2000 video stream. If this is not possible because equipment prices are high and budgets are limited, file formats such as MOV or AVI can be easier alternatives. Many collecting institutions use the proprietary codecs from companies such as AJA Video Systems and Blackmagic Design, which provide the possibility of using AVI. If the collecting institution opts to use AVI and MOV containers, it should keep an eye on developments in the area and pick the video and audio codecs to be used within the container format carefully. For example, it could make use of uncompressed PCM audio. When working with video, one should be aware that it is not only necessary to preserve the container, but also the codecs within. This should be considered when migrating or preserving the video content in another way.

8.5 Presentation and Web use

Some of the recommended file formats can be used for presentation and web use and not only for preservation, combining the use of one file format, and making it an easier task for maintaining a repository. But an institution might need or already have other file formats, audio, video and streams using specific codecs that were not made with preservation in mind but for presentation or Web use instead (See *Appendix 2* for suggestions). These file formats are not mentioned here because they don't have the recommended properties for preservation purposes, but this doesn't mean they can't be kept for other purposes. For example, the popular MPEG formats for video and audio (MPEG1, 2, 3) are not on these lists because they are all based on lossy compression algorithms, and lossy compression is not recommended as best option for preservation purposes. But for Web or presentation purposes, they are often more useful because the file sizes are smaller and can be played back or viewed in many programmes and on all platforms. Other file types, such as vector graphics and other more complex content, are not mentioned here because they are not part of our range of interest.

If the collecting institution has a lot of content for preservation in 'unsuitable' formats, it should consider developing a plan to normalise or migrate its different file formats alternatively. If not, the content might be at risk of becoming obsolete, inaccessible when requested, incompatible with certain programmes or only available in bad quality. Of course, in some cases it can seem like a lot of work keeping track of both preservation and presentation versions, and there are situations where it is unnecessary, but also times when it is needed. For example an enormous collection of images can easily be preserved in only one, good lossless format for both preservation and presentation purposes, in order to reduce the size of the repository significantly (for example jpg2000). On the other hand, with a collection of digitised videos a presentation form needs to be somewhat altered, with a frame that covers the distorted edges that an analogue video film has (for example mpeg2). This edge couldn't be seen on old TV formats, because the technology and framing was different from today's flat screens. The distortion however will usually be interesting information for the preservation copy and a different format with less compression will be preferred in order to have the best quality. Typically, with technology developing at a fast pace, the type of presentation and web formats are becoming better and better, and calling for higher quality data, so by keeping the best possible versions, one doesn't risk having to digitise all over again.

For suggestions of other file formats that are more suitable for Web or presentation purposes, see *Appendix 2*.

9. Methods for preservation

Once the file format has been chosen, the collection digitised and the data has been stored for a while, it will become necessary to consider the logical preservation and accessibility of the content. Due to technological developments, both hardware and software based, it will usually be necessary to transfer or transform digital content in order to make it compatible and usable on new platforms.

9.1 Migration

Migration is one of the most widely used approaches to preservation, and the method focuses on the file itself, not the environment in which the file is rendered. Migration includes converting one file format to another, upgrading to a newer version of the file format or moving data from one hardware type to another. (Task Force on Digital Archiving 1996).

There can be several reasons for migration, as seen below (Lawrence 2000):

1. The hardware needs to be changed before it risks breaking down due to a technical failure.
2. The software has changed, and the file format is no longer compatible with the new type of software. This could happen when there's a new operating system or other software upgrade is introduced.
3. The file format itself might be at risk of becoming obsolete because it is no longer supported or it is hardware independent. It could also just need to change because the file format is not supported by the rule of ingest and therefore must be changed to a different format.
4. For derivative copies, a preservation copy can be in one file format but the user copy of the same content can be in another to make it more suitable for presentation or web applications. For example, the master file is stored as a TIFF file but from this TIFF file a PDF derivative is made for easy access and distribution.
5. Administering the file formats has become costly and complicated due to numerous copies and versions. Hence streamlining or, as described below, normalisation is required.
6. The larger the collection becomes, the greater the requirements for organised metadata. This might result in adopting a new file format with embedded metadata properties.

Issues/Risks

Changing to new formats can come at a risk, especially if migration is the solution that is repeatedly chosen. Most common errors, which are noticeable, are a different font, layout or image colour compared to the original or even a loss of unique features that are not supported by the new format. Other issues that might occur are e.g., misplaced data or misspelled links that result in misplaced metadata or links that no longer function. Truncation errors occur typically in text document when data is migrated into a fixed size that doesn't fit. The result of this is, for example, that a text no longer fits into the table it was meant for (Gerrard 2004). More complex data, such as audio-visual material, can entail more complicated issues as well. For example, sometimes it is not the file container itself that needs conversion but the video and audio stream embedded within the container and the type of codec used to stream them. Therefore it's important to consider which type of codec is needed when the container is migrated. If not, the risk is that the compression or decompression of the video or the audio part becomes impossible, because the codec is not supported anymore.

Most of these errors can be avoided with good planning, testing, monitoring and automation. Migration is still one of the easiest ways of preserving specific data. Since there are many ways to migrate, the main issue is how to keep the content authentic and true to the original when migration continues to be the chosen strategy for preservation.

Since each migration cycle might require different procedures and equipment, the cost of migration can be difficult to predict (Lawrence 2000). The number of staff needed and the knowledge required of them might vary in time.

Recommendations for migration procedures

The types of digital objects suitable for migration are text documents, spreadsheets, presentations, audio, video and images. It is recommended to use a purposeful converter, preferably an open source one that produces the least amount of errors. All files should be checked before and after conversion for readability and errors. A simple way of achieving this is by opening and playing the content in suitable software and comparing properties such as dimensions or length of film, depending on the type of file. The embedding of metadata about the migration process should be done during conversion. It should be done both in the new file format itself and in the metadata database. Complex, software-based art, net art, or e.g., art made for CD-ROMs will have extra requirements to guarantee the preservation of their significant properties. When the files of software based art is stored on obsolete equipment or in an obsolete format, migration might affect the functionality. It cannot therefore be used as the only solution. A more complex procedure, such as emulation, would be needed.

9.2 Migration - Normalisation

In order to avoid storing multiple formats of the same material, with the risk of some of them becoming obsolete, it is possible to normalise them. This means picking only a few well-known standards and converting all odd formats into the recommended formats. It is a way to streamline the collection. As part of normalisation the files' metadata can be systematically ordered and should contain consistent information (Russell 2008). This will facilitate the ingest process and future migrations, since fewer conversion programmes are needed. Typically, in larger institutions, there are strict rules for what type of data a producer can deliver to the institution, so the normalisation is already solved before the data is ingested into the repository.

9.3 Technological or hardware preservation

The preservation of hardware for future accessibility of files dependent on hardware or obsolete platforms might be necessary for some art works. Technological preservation means that a given institution collects and maintains a certain type of technology in order to access, play or view specific data. This could for example be the case when a videotape format and the corresponding type of tape player are no longer being manufactured. To make sure that the institution's own equipment, operating systems and software last as long as possible, it is recommended to collect spare parts or extras as a solution when something stops working. But in the long run this will not solve the problem because it does not integrate the data into new platforms and in some cases also technicians with the right expertise will disappear. For a short term, technological preservation could be a solution until it is time to migrate or use some other method to transfer the data from one hardware and/or carrier to another. Sometimes technical preservation will be necessary as part of the long-term solution because the digitisation of obsolete analogue or digital tape formats is not possible without the obsolete playback equipment.

9.4 Emulation

Contrary to migration, which focuses on the specific file, emulation is a method for the preservation of the original environment in which the file can be rendered (Lawrence 2000). Emulators are programmes that recreate computer hardware and software. This approach is often useful for complex data (e.g., net based art, computer games, or other software dependent material). By recreating the initial environment the data can be shown on new software and platforms where the old software would normally be out-dated. However the newer the technology, the more complicated emulation gets. Nowadays emulation is the only option when software or an operating system is obsolete. Preferably an emulator should be made before the original version

is completely obsolete and inaccessible. Without any references to how the old system was formed, an emulator can end up functioning slightly differently from the original.

The creation of an emulator

The components needed to create an emulator (Lawrence 2000) are the data to be preserved, but also information about:

- The application software that was used to generate the file.
- The operating system that the application functioned in.
- The attributes of the hardware environment in which the software was rendered.

An emulator can be built to work on three levels of complexity:

- The programme level.
- The system software or operation system level.
- The hardware level, where the hardware is emulated - usually in a virtual way (Verde gem et al. 2006).

Some examples of these levels are:

- The programme level: WordPerfect emulators, computer games.
- The system software or operation system level: MS-DOS emulators for Windows 95/98 (Lawrence 2000).
- The hardware level: Commodore 64 joystick emulators (e.g., for the iPhone), daemon tools functioning as a CD drive which is not physically there, in order to get programmes onto a computer without a DVD drive.

Issues/risks

The biggest challenge with emulators is that they also need to be preserved in order to be useful on future systems. All the components of the emulator need to be readable, usable and well documented. If you have a completely obsolete digital object and you don't have any references, the emulator becomes difficult and more costly to make. There are different ways of preserving emulators described by Verdegem et al 2006:

1. Chaining (layers of emulators): The first emulator is preserved. However, to make sure it works on a possible new platform, another emulator is made for the new platform whereas the old one works on the inside. In such a way one can obtain a new emulator, which reflects the old platform and makes it possible for the old emulator to work on it. This might be a solution at first. However, the more layers of emulators there are the more complicated it becomes to avoid unstable and unsupported functionalities;
2. Emulator migration: Another approach is to transform the emulator to make it work on the new platform. This conversion makes sure that there is still only one emulator doing the job first intended. This approach has been tested in some instances, but it's still not defined whether the emulator will remain accessible in the long run. The risk is the same as with the migration of files: the more times one migrates, the greater the possibility of finding errors or marked changes in the originality of the programme and its rendering of files;
3. Modular emulation: Instead of an end-to-end hardware emulator, all components of the hardware are made of smaller emulators. Each of these emulators reconstructs different parts, such as the CPU, hard disc, graphic card, memory and so on. This approach requires a database that keeps track of the emulators and can retrieve metadata about them. Although this is complex, it might end up being an easier preservation solution than the others because one doesn't have to change all emulators every time a new platform is used.

All these approaches are still quite new and research on the longevity of emulators is still necessary. However, the preservation of emulators might be a good alternative for migration if it becomes increasingly more stable and easy. In 2005 The National Library of the Netherlands

made a testbed calculation of the costs of hardware emulation versus migrating data, such as text documents, spreadsheets and databases. The conclusion was that in the short term it would be more costly to build and maintain a hardware emulator, but over a period of 7-15 years it would be more beneficial (Testbed 2005, p.20-23). The benefit is that only one emulator needs to be developed for all records created on one platform type. This is contrary to a migration approach that would call for different applications for each type of file format. One should keep in mind that the Testbed might not take into account how the technology and the cost of equipment will change over time. Also the fact that technology is becoming increasingly more complex will make it more and more challenging to build emulators.

9.5 Emulation – Virtualisation

The idea with virtualisation is to have a bridge between the original software, or an emulated version of the software, and any future platform. Simply put virtualisation usually consists of making a basic processor and memory, written in a natural human-readable language but also in a machine-readable language (Gladney et al. 2005). In a sense virtualisation is just a very simple and basic emulator, which serves as an interpreter or translator of old to new. Virtualisation should make it possible to interpret the universal language in which a given software is made (the bit stream) and translate it into the language of a new platform or operative system (Gladney et al. 2005). The difference between an emulator and virtualisation seems to be a bit blurry. On the one hand an emulator is basically a fully functional entity (programme, hardware component or operating system) that is made to work in specific settings. On the other hand virtualisation is not a fully functioning entity but more a bridge or a part that can make old data work on new. However, an emulator is often needed in combination with virtualisation.

9.6 Encapsulation

This method is a bit like the AIP described in the OAIS model. Encapsulation is when the digital object is stored with all necessary components and information needed to preserve and render it in the future. This includes an operating system, the digital object itself, original processing software, and documentation of emulator specifications for hardware and metadata. It also labels the content and describes the sequence of events needed for rendering the object (Rothenberg 1999). This should make the digital object platform independent so it can be stored for a long time (although how long is not specifically stated, which of course is dependent on the content).

Encapsulation might have to be done in combination with virtualisation. When the original hardware platform is for example no longer available, it could be a solution to use a more contemporary hardware platform with a host operating system for the encapsulated guest operating system. Such a process would require making a virtualisation.

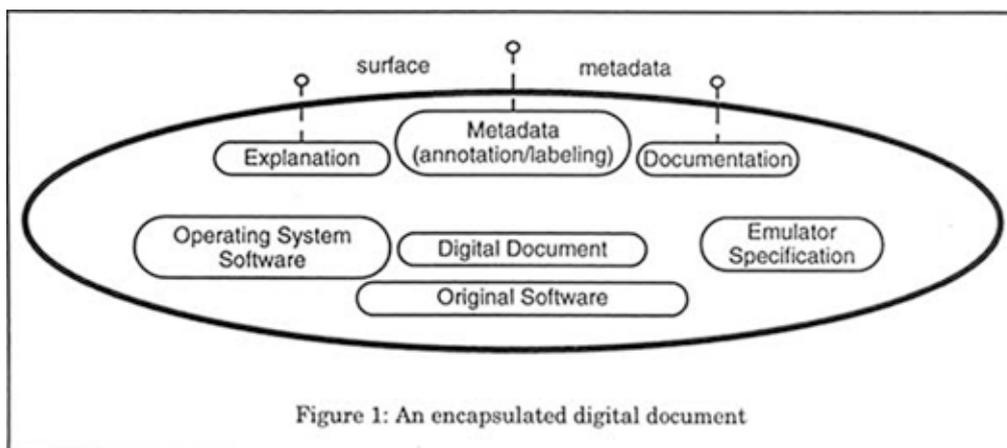


Figure 1: An encapsulated digital document

Figure 3. Rothenberg 1999

Issues/Risks

Lee et al (2002) describes encapsulation as:

“Encapsulation can be considered to be a type of migration technique. Although documentation may delay the need for migration for a long time, the encapsulated information will eventually need to be migrated. Therefore, encapsulation techniques can be applied to the digital resources the format of which is well known and that are unlikely to be actively accessed “

Encapsulation is not recommended for digital objects in use. The encapsulated object should only be for storing information that does not need to be changed for a long time. The formats and descriptions are chosen carefully so that they can be stored for a long time. But the risk that Lee et al imply is that it can be difficult to maintain systems that can actually read the encapsulated items for very long periods. Knowledge about the format must be preserved outside the encapsulation and migrations might be necessary in order for the encapsulated information to have any worth. Encapsulation could then be a method for keeping the authenticity of the files intact longer than frequent migrations might do.

9.7 Cloud Computing

Cloud computing stands somewhat apart from the other methods described in this section, because it is mostly about storage and access and not so much about long-term preservation. The files as such would still have to be migrated to new formats once in a while. The advantage of the cloud comes with the possibility of more stable and inexpensive storage. Cloud computing is an emerging technology for easy access and storage of files. Wang and Laszewski (2008) define the computing cloud as:

“...a set of network enabled services, providing scalable, QoS guaranteed (Quality of Service), normally personalized, inexpensive computing platforms on demand, which could be accessed in a simple and pervasive way”.

There are a lot of different web applications, such as IBMSmartCloud (IBM 2011), Amazon Elastic Compute Cloud (Amazon 2011) and Microsoft Cloud Power (Microsoft 2011) to name just a few. Google has even launched a complete operating system based on their chrome browser, Google Chrome OS (Levy 2011). Everything is stored in ‘the cloud’ and therefore accessible from any computer. This makes it easier to share and edit the same documents (Buyya et al 2008). It also eliminates the danger of having the server in one place, since the cloud hosts servers in many different places. Because the data is stored on the web, it should be easier to recover any lost data through the company who supplies the programmes.

To give a basic idea on how cloud computing could be utilised for digital preservation, here are some simple technical specifications (Robert de Geus, personal communication 2011):

1. The cloud should store files based on the rules of the defined bit preservation, described earlier in this article;
2. The cloud should implement the CAP theorem (Gilbert, S. et al) that stands for:
 - Consistency (all nodes²³ see the same data at the same time);
 - Availability (node failures do not prevent survivors from continuing to operate);
 - Partition tolerance (the system continues to operate despite arbitrary message loss).
3. By replacing or adding nodes the cloud will reorder itself to conform the bit preservation rules and the CAP theorem;
4. The cloud should implement a predefined consistency safety, which describes how many nodes can fail before irreparable data loss occurs.

²³ A node is defined as a connection point in a network, either a redistribution point or an end point for data transmissions. In general, a node has programmed or engineered capability to recognise and process or forward transmissions to other nodes.

If these preparations are taken into account, cloud computing should result in a better management of digital data, cheaper administration, better availability, and higher preservation standards than existing technologies.

Issues/Risks

Cloud computing is still a fairly new technology, coined in 2007 (Wang 2008) and therefore still poses risks of data loss, contrary to what the companies might say. At present, research is needed in order to know if the cloud is a good tool for digital preservation because cloud computing as such does not work as a preservation precaution alone. The bit stream still has to be saved somewhere and still has to reside on one or more physical devices, which will cost extra money. Also there are still a lot of questions regarding privacy and ownership rights and whether such rights are being protected.

9.8 How to choose

All methods or preservation strategies mentioned in this section have pros and cons. They work for different types of digital information, and the choice of strategy varies depending on what type of end result we wish to obtain. A simple model below shows the basics for choosing a specific method (Figure 4). It is partially inspired by the National Library of New Zealand (NLNZ) (2003) and a model that Sofie L. Henriksen developed for a bachelor project on the subject. It starts with the assessment of a given digital repository. Depending on what type of content the collecting institution has and what state the content is in (obsolete, in use etc.) there are different solutions suggested by following the arrows. For example, if one has images in an obsolete format that can no longer be migrated, the use of an emulator might be a solution. If one has images that are not obsolete, but one actively uses them and one needs to change the software or platform, a migration process will be a good option.

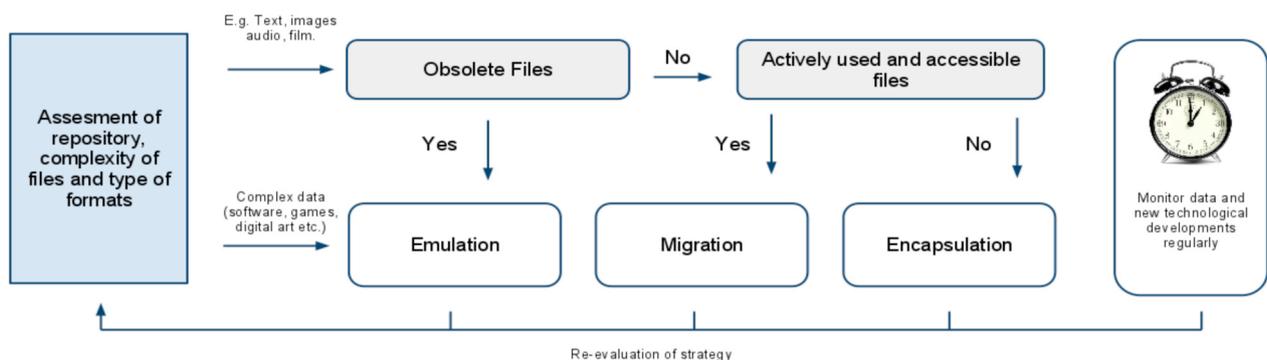


Figure 4. Model for choosing the right preservation method, depending on type of data and its status.

10. Preservation planning

One thing is picking the right method, but in order to choose the right file format, method, equipment, software and time to implement it, it is crucial to have a preservation plan that can help highlight the pros and cons. Hans Hofman from the Netherlands National Archives defines a preservation plan as the following:

*“A **preservation plan** defines a series of preservation actions to be taken by a responsible institution to address an identified risk for a given set of digital objects or records (called collection).”*

A preservation plan can be made from scratch, but for large repositories or collections of very mixed content, there is a tool to help. The European project group Planets has made a preservation-planning tool called PLATO (www.ifs.tuwien.ac.at/dp/plato/intro.html), which helps define and test the content and the suitable solutions at hand. The workflow follows three phases with eleven steps that all give numerical values. The end result and suggestions are based on these values and are not subjectively chosen (Figure 5). This can be a long process and difficult to use. However, if one has the time and knows how to define what one wants, it can be a good way of figuring out what conversion tools and file formats to use. The preservation plan will also be fully documented afterwards. The steps explained below can be used for a self-made plan as well.

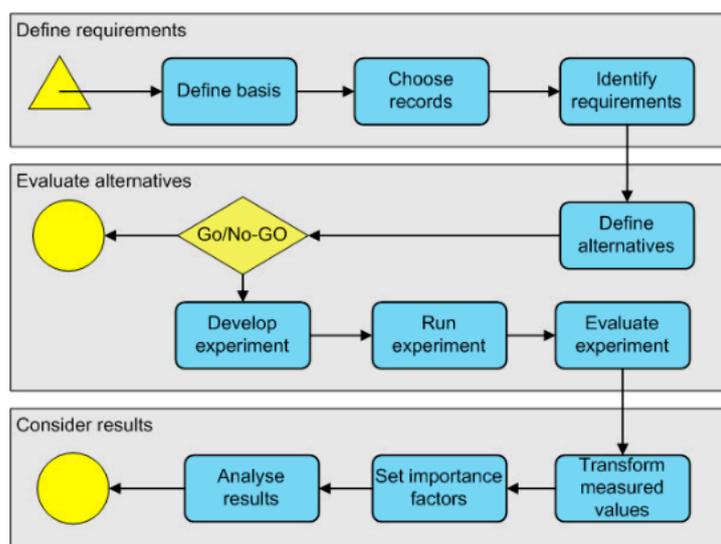


Figure 5. PLATO steps. Stroll, S. et al 2007

Based on the PLATO User Manual:

1. **Define Basis:** One starts by defining the repositories' content, such as what type of file formats are in the collection, how many there are and when and how they were generated. The institution's overall policies for preservation should also be taken into account.
2. **Choose Records:** A few examples of files are picked for testing the following methods. The example files have to represent the variety of the collection.
3. **Identify Requirements:** By identifying requirements the plan has to answer questions like: What is the content going to be used for? How long is it going to be stored and how do we want to

access the information? What information is the most important? This is a fairly complex procedure and all stakeholders should be involved (curators, technical staff, end-users) in order to cover all aspects. PLATO provides a tree structure, ordering topics according to their importance.

4. **Define alternatives:** Define which tools and software can be used. This includes all possible file formats and codecs, as well as hardware on which to store the files and software to access, edit and retrieve the data. More than one possibility should be found.
5. **Go/No – Go:** By theoretically and realistically looking through all alternatives of point 4, and trying out what can be easily tested, it is possible to find the methods that can live up to requirements and goals.
6. **Develop Experiments:** Pick the best method or methods from point 5 and make a full experiment using the example files chosen at the beginning.
7. **Run Experiment:** Run the experiment entirely, digitise, ingest, set up automatic validation of files, create preservation metadata, and store it on the chosen hardware and so forth.
8. **Evaluate Experiments:** Evaluate the results and see if they live up to requirements and goals.
9. **Transform measure values:** Measurements taken during the experiments can be scaled differently. In order to compare different methods, they can be transformed to a uniform scale using transformation tables that PLATO provides. This gives all results a quantifiable value. If some things don't quite measure and score too low, there is a possibility to tweak or change parts of the procedure to see if this will improve the method.
10. **Set important factors:** Find points in the plan that are important or crucial for further development. In this way one can focus on the important parts first, such as the file type chosen for digitisation, or the validation of files after digitisation.
11. **Analyse results:** By using the tables PLATO provides an overview of which methods works best. If a method is found suitable, one has made a digital preservation plan. Every step should be documented and be part of the collecting institution's preservation or digitisation policy.

The process breaks down the requirements into measurable criteria, ending with an objective tree defining what is wanted and needed (Becker et al. 2008). The tool is available on the Internet and guides the planner through the process. If you would like to know more about it, check PLATO's user manual (see PLATO, under references).

There are several examples on Planet's homepage of how PLATO has been used as a tool to work out the best preservation goals and how to get there. Larger libraries have used it for dealing with big collections of scanned images. A thesis has also been written on the use of PLATO for the preservation of a videogame. To get an idea of situations and solutions we recommend reading Christoph Becker & Andreas Raubers' article *Four cases, three solutions: Preservation plans for images* (2011). They describe four different collections of scanned books, newspapers and images and what solutions were chosen with the help of PLATO, depending on the criteria described above. For example the Royal Library of Denmark used the PLATO tool to figure out which files would be necessary during a digitisation project of a big photographic negative collection. It was suggested that they choose two different quality image files for two different purposes: to digitise a collection of negatives in two qualities. One copy was digitised in high quality (1800ppi, 16bit, RGB) because the scan had to replace the original, very degraded, analogue version. To save storage space, the other part was scanned in a lower quality (1800ppi, 16bit, greyscale) because the scans were only to be used as a digital copy, while the analogue

original still remained available (Becker et al 2011). If a better version of the digital copies was needed, the original could be scanned again. Of course it is necessary to make sure that the plan is part of the collecting institution's overall policy and that staff members are available to actually do the tasks at hand or at least be the contact person between external vendors and the institution. By documenting every step and decision, it will always be known what was actually done which makes it easier to adjust the plan to fit future preservation plans.

11. Preservation Metadata

Metadata is information about information, or data about data. Basically, it is structured information that describes, locates, or explains an information resource in order to make it easier to use, retrieve, and manage (NISO 2004). Usually this information is put into a schema, which is a kind of metadata format. No matter what type of metadata schema is used, there are always some practical properties recommended when choosing one.

11.1 Recommended properties of metadata

Open source language	Standard format scheme	Structured and relational metadata	Extensibility
Should be written in a standard, well-known language, such as XML.	The chosen metadata schemas should be suitable for the content and an acclaimed open standard.	<ul style="list-style-type: none"> - Well-defined sequence of information. - Possibility for linking to a separate database. - Essential metadata should be embedded within files. 	Standard schema that can translate or render a range of other metadata formats.

Table 7

There are several categories of metadata, such as descriptive, structural, administrative and technical. It is important that a given digital object has all kinds of metadata, but in this deliverable *D6.1* the focus will be on the type of metadata that's important for preservation purposes. Preservation metadata is intended to support and facilitate the long-term retention of digital information (OCLC 2001).

According to the OCLC working group (Online Computer Library Center), preservation metadata should be used to:

1. Store technical information supporting preservation decisions.
2. Document preservation actions taken, such as migration or emulation.
3. Record the effects of preservation strategies to ensure the authenticity of digital resources over time.
4. Enable objects, for which the library has assumed preservation responsibility, to be identified.

If this information is safeguarded, it should be sufficient to support knowledge and insight into how to preserve the bits as well as the access and logical sense over the data for a long time (OCLC 2001).

11.2 What is preservation metadata?

Preservation metadata is usually a combination of descriptive, structural, technical and administrative metadata (Dappert et al 2010 and Caplan 2009):

Descriptive metadata: Descriptions of what digital object it is, when and by whom it was made, it's name and where to find it.

Structural metadata: Logical and structural relationships. It can be the page order in a book or the information about which images belong to which websites.

Technical metadata: Gives information about file type, which hardware and software was used when creating the object, and which platform or software will be needed to render it. It can also be

checksums and other authenticity verifying information. Depending on the type of file, technical metadata should also include things such as image size (images) or frame rate (video), this part will be elaborated in the following tables. A checksum should be included too, in order to validate the integrity of the data. The technical metadata is usually incorporated into the digital preservation plan because it provides information on what is needed to access and render the object correctly.

Administrative metadata: This includes rights managements and documented preservation activities. Any change or modification to the object should be included in administrative metadata. These changes should also be described in the metadata database. The metadata could for example indicate that the file has been migrated.

Recommendations for what to include

Recommendations for which elements to include in preservation metadata can be seen below, based on the NLNZ 2003 appendix:

Object	Process	File
Name of object	Object identifier	Object identifier
Reference number	Process type	File identifier
Object identifier	Purpose	File path
Group Identifier	Person/agency performing process	Filename and extension
Persistent Identifier	Permission	Former filename
Preservation Master Creation Date	Permission date	File size
Logical Composition	Hardware used	File date and time
Structural Type	Software used	MIME type (e-mail or web standards)
Comments	Steps	File format
Hardware environment	Results	File format version
Software environment	Guidelines	Target indicator
Installation requirements	Completion date and time	
Access inhibitors (for example any encryptions)	Comments	

Table 8

Depending on file type, specific technical metadata are also needed (based on NLNZ 2003):

Audio	Text	Images	Video
Resolution ²⁴	Character set ²⁵	Resolution ²⁶	Frame dimensions ²⁷
Duration ²⁸	Mark-up language ²⁹	Dimensions ³⁰	Duration ³¹
Audio bit depth ³²		Bit depth ³³	Frame rate ³⁴
Audio bit rate ³⁵		Colour space ³⁶	Codec method ³⁷
Compression ³⁸		Colour management ³⁹	Aspect ratio ⁴⁰
Encapsulation ⁴¹		Colour map reference ⁴²	Scan mode ⁴³
Channels ⁴⁴		Orientation ⁴⁵	Sound indicator ⁴⁶
		Compression ⁴⁷	Video bit rate ⁴⁸
			Video bit depth ⁴⁹

24 The rate of sampling, in samples per second, used to create the audio file. Also known as sample rate or sample frequency. E.g., 32.100 Hz, 44.100 Hz, 19.2000 Hz.

25 The character set used when creating the file. E.g., ASCII; Unicode; EBCDIC, UTF-8.

26 The spatial resolution of the image, expressed as pixels per inch or cm (ppi, p/cm) or dots per inch or cm (dpi, d/cm). E.g., 600 dpi; 320 dpi, 1500 d/cm

27 The resolution in pixels of a single still frame. E.g., 640 pixels x 480 pixels

28 The length of the audio recording in hours, minutes and seconds and three digits for representing decimal fractions of a second. E.g., 01:27:38:247.

29 The type of mark-up language used to mark-up the document. E.g. SGML, XML, HTML

30 The dimensions of the image, expressed as the number of pixels along the vertical and horizontal. E.g., 4096 x 6144 pixels

31 The length of the video recording in minutes and seconds, or minutes, seconds, 100ths of seconds. E.g., 01:27:38:247

32 The word length used to encode the audio. Consequently an indication of dynamic range. It is the maximum number of significant bits for the value without compression. E.g., 16, 20, 24 bits.

33 The number of bits per component for each pixel. e.g. 1 = 1 bit (bitonal); 4 = 4 bit greyscale; 8 = 8 bit greyscale or palletised colour; 8,8,8 = RGB; 16,16,16 = TIFF, HDR (high dynamic range); 8,8,8,8 = CMYK.

34 The rate at which the video should be shown to achieve the intended effect – expressed in frames per second (fps). E.g., 25

35 In audio bitrate represents the amount of information, or detail, that is stored per unit of time of a recording. It indicates the amount of compression used. E.g., 8 kbit/s, 256 kbit/s, 1,411.2 kbit/s.

36 Designates the colour space of the decompressed image data. E.g., 0, 1, 2, 3, 4, 5, 6, 7, 8.

37 The name, including version level, of the codec method applied to the video. Note that video compression, or bit rate reduction is a non-reversible 'lossy' process. E.g., DivX 5.0.5.

38 The name of the compression scheme, noise reduction scheme, or other non-linear processing applied to an audio signal. Note that audio compression, or bit rate reduction is a non-reversible, 'lossy' process. E.g., MPEG 3, Dolby A.

39 The name of the International Color Consortium (ICC) profile used. E.g., PhotoCD; OptiCal; Profile/80; Softproof (Photoshop plug-in).

40 The desired aspect ratio of the image on screen. E.g., 4:3.

41 The name and version level of the delivery format of the file. E.g., Real Audio II.

42 The location of the file containing the colour map. E.g., [URL]

43 An indicator showing whether the digital item is scanned in a progressive or interlaced mode. E.g., Progressive, Interlaced.

44 A classification of the sound format type identifying the number of channels and how they are related to each other. E.g., mono 2 channel stereo 5 channel surround, other.

45 Orientation of the image saved on disk e.g. normal, normal rotated 180°. E.g., 1 = normal*; 3 = normal rotated 180°; 6 = normal rotated cw 90°; 8 = normal rotated ccw 90°; 9 = unknown.

46 An indicator of the presence of sound in the video file. E.g., Yes; No. Note: If the value is 'yes', then the video file will also be associated with an instance of the Audio metadata in addition to the Video metadata.

47 The type and level of compression. E.g., 4 = ITU Group 4.

48 In video bitrate represents the amount of information, or detail, that is stored per unit of time of a recording. It indicates the amount of compression used. E.g., 16 kbit/s; 1.5 Mbit/s; 3.5 Mbit/s; 9.8 Mbit/s; 25 Mbit/s; 29.4 Mbit/s; 40 Mbit/s.

49 The word length used to encode the video. Consequently an indication of image quality. It is the maximum number of significant bits for the value without compression. E.g., 8; 10 bits.

			Pixel aspect ratio ⁵⁰
			Field order ⁵¹
			Colour space ⁵²
			Chroma sub sampling ⁵³

Table 9

Often metadata formats are made to suit specific types of information and one metadata format does not necessarily fit every type of file in a repository (DCMI). Some are specifically made for libraries, for example MARC; others are made to describe scientific datasets or art works.

The ideal situation calls for a metadata database connected to the individual files, as well as embedded metadata within the files (Verheul 2006). This is in order to preserve the information as well as possible. By having the essential, descriptive and non-changing metadata within the file (e.g., name of artwork, creator, a date and a checksum), the file can always be identified no matter where it has been put. However, it is unnecessary to have new or changing information embedded in the file (e.g., administrative metadata). Embedding this kind of information would take too much time and it would be hard to keep track of each single file. When changes are made (e.g., when a file is copied to another file format or the ownership changes), this should be registered in the metadata database instead. When the file is lost or becomes obsolete, we can still find information about the file elsewhere and every single file does not have to be retrieved to add new information. If it is not feasible to have metadata in both locations, the priority should be to have a database of the essential metadata that is continuously maintained and linked to the files. The larger the collection is, the more it is recommended to consider implementing embedded metadata.

11.3 Recommendation of preservation metadata standards

The preservation metadata can be incorporated with other metadata in a system (like a database). This metadata is usually called a format or a schema, and there are many different types developed for different purposes. There are also quite a few systems for collecting metadata schemas. Below are some recommended schemas for preservation metadata.

PREMIS

In 2001 a group specifically addressing preservation metadata was formed called PREMIS (PREservation Metadata Implementation Strategies). PREMIS is a set of strategies and elements described and developed for encoding and storing preservation metadata in a digital preservation system (NISO & OCLC/RLG), based on the OAIS model. PREMIS suggests having metadata embedded within the file, as well as in a database. For the schemas, they recommend using the metadata container standard METS (Metadata and Encoding Transmission Standard) for making a database together with the PREMIS schemas. PREMIS has a lot of documents and introductions on the use of METS as a database for different types of metadata formats. PREMIS and METS schemas are based on XML, which is good for creating and storing preservation metadata, and there is already a set of preservation specific vocabulary.

⁵⁰ A mathematical ratio that describes how the width of a pixel in a digital image compares to the height of that pixel. E.g., 1,09 (59:54); 1,09 (12:11); 1,46 (118:81); 1,45 (16:11); 0,90 (10:11); 1,21 (40:33); 1,32 (4:3).

⁵¹ In video, a field is one of the many still images that are displayed sequentially to create the impression of motion on the screen. Two fields comprise one video frame. When the fields are displayed on a video monitor they are 'interlaced' so that the content of one field will be used on all of the odd-numbered lines on the screen and the other field will be displayed on the even lines. It is crucial to know which of the two fields is 'dominant'. E.g. odd; even.

⁵² Designates the colour space of the video data. It is an abstract mathematical model describing the way colours can be represented. E.g. YCrCb; XYZ.

⁵³ Practice of encoding video images by implementing less resolution for chroma information than for luma information, taking advantage of the human visual system's lower acuity for colour differences than for luminance. E.g. 4:1:1; 4:2:0; 4:2:2; 4:4:4.

To implement and use the PREMIS vocabulary in a XML based database it is necessary to have some knowledge of how to write in XML. It is possible to find an example of how a metadata XML schema looks and fill it in with the metadata needed for a specific case (example in Appendix 4). To learn more about implementing metadata in general, see the deliverable *D3.1 Metadata implementation guidelines for digital contemporary artworks*. For more information about PREMIS, see <http://www.loc.gov/standards/premis/tutorials.html>

OWL

PREMIS OWL, the ontology to PREMIS, has recently been published in collaboration with the Library of Congress and IBBT (leader of DCA WP3). It is a metadata schema in RDF instead of XML, which can be an advantage when working across different databases. Pricilla Caplan described OWL as:

"This OWL ontology allows one to express the same information in RDF (Resource Description Framework). With this alternative serialization, information can be more easily interconnected, especially between different repository databases. Information in RDF can be also easily and flexibly queried, which can be an interesting option for the data management function of a repository. The PREMIS OWL ontology also reaches out to preservation-specific vocabularies already published by the Library of Congress."

Pricilla Caplan (2011)

OWL is not meant to be a complete replacement for PREMIS. It is meant to be an addition so that it is more convenient to work across different databases, which was complicated to do beforehand.

Other

When setting up a metadata database, it would be logical to include other types of metadata to improve the searchability and retrievability of digital objects. For this, there are several types of formats. Some specifically recommended are MARC/MARC21 (Machine Readable Cataloguing) and MODS (Metadata Object Description Schema). Both schemas are originally developed by Network Development MARC Standards Office for libraries, but can also be used in other contexts.

It is important, as with file formats, that the metadata is compatible with other types of metadata. Some participants in the DCA project might have used metadata schemas such as CDWA, CDWA lite, EN 15744 and EN 15907. These schemas are not recommended for recording preservation metadata because they are not very widespread, don't have a specific preservation vocabulary and some of them have not been updated since 2000 (CDWA). This might cause problems when merging different metadata schemas together and result in information changing form or not being adaptable with newer systems.

If you would like to know more about metadata for artworks, see *D3.1 Metadata implementation guidelines for digital contemporary artworks*.

11.4 Extraction and creation of metadata

A metadata database is built up of information extracted from files, plus additional information. If the metadata embedded in the files is not complete (refer to lists above), the missing information has to be added to the schema. To give an example of what a metadata schema looks like, see appendix 4. This example shows metadata in an XML schema. It uses MODS, METS and PREMIS metadata all in one.

The metadata to be introduced into the schema can be taken from a variety of sources (OCLC/RLG):

- Some is ingested into the archive storage with the object.
- Some is created by the archive storage as a result of internal processes.

- Some values are not stored in the digital archive but can be supplied on output.

In the schema, basic information can be written and kept. To keep track of the schemas a database is usually needed. To extract metadata information from files, validate the file's authenticity. To create metadata for the metadata database, it is necessary to use some tools. Sometimes it is also useful to know which type and version of format the file is. Specifically when working with multimedia formats, it might be necessary to know the codecs that are embedded within the container. In this case it is not enough to have the file type extension (.DOC, .PDF etc.). There are many tools that can be used for the validation and creation of metadata on file formats and codecs. Usually the same tools can be used for metadata extraction. The recommended tools are briefly described below.

JHOVE (Jstor/Harvard Object Validation Environment) provides functions to perform format specific identification, validation, and characterisation of digital objects. JHOVE provides robust and detailed metadata that can be implemented directly into PREMIS XML schema elements. JHOVE is good for a small set of standard-based file formats but it requires a high level of technical skills to install and use JHOVE supports most image formats, but not video codecs. To install the program go to: <http://hul.harvard.edu/jhove/distribution.html>. The installation and use of JHOVE is not a straightforward process. A step-by-step guide might be required. The Belgian website www.projectceest.be/index.php/Handleiding_JHOVE is a good source. Unfortunately it is only available in Dutch, but with a Google translation and a little patience it can be really helpful.

In April 2011 a new version of JHOVE was launched: **JHOVE2**. JHOVE supports a broad range of formats. JHOVE2 does not support some of the ones that JHOVE does (AIFF, GIF, JPEG, HTML). A novelty is the validation of ICC colour profiles. While JHOVE has a graphic user interface, JHOVE2 only has a command line interface. JHOVE2 can be downloaded from <https://bitbucket.org/jhove2/main/wiki/Home>. More information on JHOVE2 can be found on https://bytebucket.org/jhove2/main/wiki/documents/Abrams_a70_pdf.pdf. The Belgian website http://www.projectceest.be/index.php/Handleiding_JHOVE2 is also a good source (only in Dutch).

DROID (Digital Record Object Identification) is developed by Pronom, a working group for the National Archives of the UK. It is an automatic file format identification tool that is open source and should work on several operating systems. DROID handles a much larger range of formats than JHOVE (also video codecs) but has a more limited metadata output. In order to use DROID metadata for a PREMIS XML schema, the information also has to be converted first. When one downloads DROID, one obtains a set of files. One of them is a text file that provides instructions for starting the programme. Depending on what platform one is using (Microsoft, Mac or Linux), different files will be needed. To install DROID, go to: <http://droid.sourceforge.net/> (The Rockefeller Archive Center, 2009). Or for more guidance, try yet another useful guide from CEST in Belgium, http://www.projectceest.be/index.php/Handleiding_DROID (again only available in Dutch).

To obtain a good comparison mechanism, it is recommended to use both programmes. Other validation tools that support both video and audio codecs are FFprobe and MediaInfo. These programmes are quite complex, and require technical knowledge or experience with implementing metadata. More tools and explanations on the different properties can be seen at the Library of Congress website about PREMIS and tools:

http://www.loc.gov/standards/premis/tools_for_premis.php. In EU Scape different tools are evaluated: <http://www.openplanetsfoundation.org/blogs/2011-09-21-evaluation-identification-tools-first-results-scape>

To avoid coding and find a quick and easy way of extracting the metadata within a file, the following small programmes are recommended.

ExifTool is an easy tool to use for manual metadata extraction. Wiel Seuskens (NIMk) has written a short manual on how to use this programme on Windows and on Mac, see *Appendix 3*. For Macs a little programme needs to be built in order to use the ExifTool more easily. It should be kept on one's desktop. Every time one wants to extract the embedded metadata of a file, one should drag it over the icon of the programme, which will then automatically generate a text file with the same name as the file, but with the file extension .TXT. This text file will be put next to the one from which the metadata has been extracted. The text file contains all available embedded metadata (see example below). The data can in principal also be used in an existing metadata database. Please note that ExifTool does not validate file formats and codecs, it just extracts the embedded metadata. If this embedded metadata has been wrongly transformed, it will provide one with the wrongly transformed metadata. To download ExifTool and find more information, see <http://www.sno.phy.queensu.ca/~phil/exiftool/>

An example of a .TXT file metadata extraction made with an Exiftool:

```
---- ExifTool ----
ExifTool Version Number: 8.67
---- System ----
File Name: 73925501.jpg
Directory: /Users/SofieLH/Pictures
File Size: 44463
File Modification Date/Time: 2011:01:29 18:37:52+01:00
File Permissions: 644
---- File ----
File Type: JPEG
MIME Type: image/jpeg
Comment: CREATOR: gd-jpeg v1.0 (using IJG JPEG v80), quality = 90.
Image Width: 477
Image Height: 636
Encoding Process: 0
Bits Per Sample: 8
Color Components: 3
Y Cb Cr Sub Sampling: 2 2
---- JFIF ----
JFIF Version: 1 1
Resolution Unit: 0
X Resolution: 1
Y Resolution: 1
---- Composite ----
Image Size: 477x636
```

12. Checksums

A collecting institution can choose to maintain its digital collection itself or hand over the responsibilities to a subcontractor. Either way, there should be someone in-house who knows how to check the files. It is necessary to ensure that the files have the wanted properties and don't undergo unforeseen changes. There are several aspects that can be checked and several tools for checking them. The preservation metadata is one example (see *Chapter 11. Preservation Metadata*). But one also needs to check whether a given file has stayed the same over time.

The creation and checking of checksums is one of the most common procedures. A checksum is a calculated value of the bits of a given file. This value can also be called a hash value. The checksum can be used to validate the data. If the bits in the file change, the checksum of the file will also change. Checksums are used to make sure there is no loss of data or corruption of the file. Checksums should be overseen regularly and automatically in order to discover file corruption.

A checksum should be generated when a digital object is made. It should be stored in a metadata database in order to compare the value with the file itself. If the file is converted into a new format or newer version of the same format, the checksum will alter and so the one kept in the metadata should also be changed. This is not an error. It is simply due to the fact that the bits that the checksum relates to, change and give another result. There are different types of checksums, which are more or less secure.

CRC (Cyclic redundancy check) is used by file systems to check file copying. CRC is not unique to a file. It can be used to check whether the file copying was done correctly, but it shouldn't be used for preservation purposes.

MD5 (Message-Digest Algorithm) provides an almost unique identifier for a file. It produces a 128-bit (16-byte) hash value (that gives 256^{16} possibilities). It is possible to create a file with the same MD5 hash value, but it is very unlikely that two 'meaningful' files have the same MD5 hash value. There are even file systems that depend on MD5.

SHA-1 (Secure Hash Algorithm 1) is more secure against retrieving original data than MD5. It is usually used for password protection and security reasons.

SHA-2 (Secure Hash Algorithm 2) gives a better protection than SHA-1 because SHA-1 can be hacked. Usually it is used for password storage.

It is reasonable to generate MD5 checksums for preservation purposes, since they need to be unique but are not made for security purposes. If one knows a bit about coding, simple commands in the command lines interface of the Windows or Mac operating systems can be used. Otherwise there are many special programmes, including ones that can also be used to extract other information from files (see *Chapter 11. Preservation metadata*). Programmes for Windows are extractfile.com, fastsum.com (only MD5 checksums) and [advanced checksum verifier](http://advancedchecksumverifier.com). For Mac the terminal can use Bash.

When checksums are generated, it is very important to put the information together with the preservation metadata in a database somewhere outside archival storage. To detect file corruption it is recommended to validate the checksum every six months by comparing the checksum of each file with the values in the database.

One should keep in mind that it is very time-consuming and unnecessary to generate and verify checksums manually. This is a procedure that should be automated by technicians. They can use BASH, Perl or PHP to code commands that automatically generate a checksum, and put it into the

right place in the database. When it is time to verify, the coding can be used to run through the files and compare their checksums with those in the database.

13. Persistent identifiers

The research project ATHENA (access to cultural heritage networks across Europe) provides the basis for this chapter. It looked into how cultural institutions use Persistent Identifiers (PIDs) (ATHENA 2011). ATHENA describes the functions of PIDs as:

Identification - Using agreed strings of alphanumeric text (identifiers) to provide access, such as a key, to information in paper-based, in-house computer and online systems. They also provide access to physical objects using attached marks or labels.

Persistence – Managing the identifier in order to maintain access.

A persistent identifier consists of a uniquely picked number or other symbol that is given to an object and that provides access to it. The object can be something that is physically labelled, a file or physical object registered in a database, or it could be a digitally born file. Either way, the object needs a unique identification code in order to be found and identified. This section focuses on digital materials and the basic content of PIDs.

The PID is connected to the metadata of the object. Therefore both the object and the metadata should have the same form of identification, unless it is a copy or documentation of the object. Digital reproductions of analogue originals should have a different identification code because they are copies. It should be stated in the metadata of the digital copy that it is derived from an analogue version of the artwork.

When talking about PIDs, it is usually when they are related to the Internet and describes how links are found, retrieved and named. There are three types of standards described by ATHENA:

1. URI (Uniform Resource Identifier): A string of characters used to identify a name or a resource on the Internet. This can be an URL or an URN, or both.
2. URL (Uniform Resource Locator): An URI that specifies where a resource is available and how to retrieve it. It identifies where the object is and how it can be accessed but it does not provide the name or information about what it is. For example www.europeana.eu is an URL. By clicking on the link or typing it in one's browser one can find where one's object is.
3. URN (Uniform Resource Name): An URI acting as persistent, location-independent resource identifier that is designed to make the mapping to other namespaces easy. An URN does not point to a location and therefore might not be resolvable. For example an International Standard Book Number (ISBN) identifies a specific book but not where to get hold of it. One can put a digitised version of this book on europeana.eu, but one can also put it somewhere else. The URN only identifies the specific file/object, not where it is located.

In order to give support and use the PIDs, there are several systems to choose from. Currently there are no PID systems specifically built for the museum sector, but ATHENA recommends:

1. PURL & Handle system (Persistent URL & Handle system): Originally made for libraries. Not an open standard.
2. DOI (Digital Object Identifier): Created by the National Information Standards Organisation (NISO), which also specifies preservation metadata. Open standard.
3. OpenURL: Also developed for libraries. Open standard. Made by OCLC.
4. ARK (Archival Resource Key): Developed for libraries and medicine archives. Open standard.

The requirements for these PID systems are:

1. As with hardware systems, the PID system has to be reliable, i.e., active and up-to-date, and should be checked regularly for errors (preferably automatic).
2. The PID system has to have a reliable and committed support system, either through a subcontract company or within the organisation itself. This authority should also be evaluated regularly, especially before implementation.
3. The PID system should be flexible, and able to handle different types of collections.
4. The PID should be interoperable, through the use of open standards, in order to share the collection with a large set of users.

When managing PIDs, there are some other points to be aware of (Bellini et al & ATHENA):

1. Make clear in which environment the PID is unique. Is it only within the organisation's own system or does the object have a globally unique PID because it is on the Internet?
2. Make sure that the PID is persistent and cannot be changed or deleted. This should be done by defining what is meant by persistent and how a user can be sure of its persistence (for example with the help of checksums).
3. There should be an indication of who can access it and from where. It should be clear whether the PID is externally available or not.
4. To save money, the PID systems in use should be open source, free of charge or very low in cost.
5. As with everything else, the use of PIDs should be part of the written Collection Management and Access Policy.

For more on persistent identifiers look up the ATHENA projects booklet on their website:
<http://www.athenaeurope.org/index.php?en/1/home>.

Glossary

AIP – Archival Information Package – A package designated for a digital archive, which includes not only digital content information but also preservation description information

DCA – Digitising Contemporary Art

DIP – Dissemination Information Package – The AIP becomes a DIP when the end-user or consumer requests a package's content information

DROID – Digital Record Object Identification – Tool for verifying files and their format

HDD – Hard Disk Drive

ISO – International Organization for Standardization

JHOVE – Jstor/Harvard Object Validation Environment – Tool for verifying files and making metadata.

LTO – Linear Tape Open (Archival tape)

MARC – Machine-Readable Cataloguing – Metadata standard

METS – Metadata Encoding and Transmission Standard – Metadata structural standard

MODS – Metadata Object Description Schema – Metadata standard

Nodes - In a network, a node is a connection point, either a redistribution point or an end point for data transmissions. In general, a node has a programmed or engineered capacity to recognise and process or forward transmissions to other nodes

OAIS – Open Archival Information System – Model for archival preservation storage and components needed for ideal preservation of digital content

OCR – Optical Character Recognition

PCM – Pulse Code Modulation is a method used to digitally represent sampled analogue signals, the higher the sample rate the better the digital sound represents the original analogue signal

PDI – Preservation Description Information – Information package needed for preserving a given digital object for the long-term

PID – Persistent Identifier, used for preserving links to digital or analogue objects

PREMIS – PREservation Metadata: Implemented Strategies – Data Dictionary developed by a group from the Library of Congress, US

RDF – Resource Description Framework

SSD – Solid State Drive

SIP – Submission Information Package – The information package made by the producer, before it is made into an AIP

XML – eXtensible Mark-up Language - Standard mark-up language, combines text and extra information about the text

References

Main literature references

The information in this guideline has been acquired from many sources. As there is a large amount, it has not been necessary to develop our own terminology or results; instead it has been a process of picking the right resource. There are references to a lot of different articles and websites, but some of the most important will be mentioned below.

General information on digital preservation:

- CCSDS (2002). Recommendation for space data system standards. Reference Model for an Open Archival Information System (OAIS) (Online).
<http://public.ccsds.org/publications/archive/650x0b1.pdf>
- Dollar, C. (2009), Digital Preservation Capability Maturity Model Version 3.

Information on file formats:

- Library of Congress, http://www.digitalpreservation.gov/formats/fdd/browse_list.shtml
- Digitaal Erfgoed Nederland (Digital Heritage Netherlands), www.den.nl
- International Organization for Standardization, www.iso.org
- CEST – Cultureel Erfgoed Standaarden Toolbox,
<http://www.projectcest.be/index.php/Standaarden>

Information on persistent identifiers:

- ATHENA WP3, McKenna, G. & Wyms, R. (2011). Persistent Identifiers (PIDs): Recommendations for Institutions; <http://www.athenaeurope.org/index.php?en/1/home>.

Information on preservation metadata:

- PREMIS (2011). Data Dictionary for Preservation Metadata version 2.1(Online).
<http://www.loc.gov/standards/premis/>.

Adler, M, Boutell, T., Bowler, J. (2003). Portable Network Graphics (PNG) Specification (2nd Edition) Information technology — Computer graphics and image processing — Portable Network Graphics (PNG): Functional specification. ISO/IEC 15948:2003 (E), edited by David Duce: Oxford Brookes University, W3C (Online). <http://www.w3.org/TR/PNG/> (April 2011).

AFP (2009). Google apologizes for Gmail breakdown. My Digital FC (Online).
<http://www.mydigitalfc.com/it/google-apologizes-gmail-breakdown-038>

Amazon (2011). Amazon Elastic Compute Cloud (Online). <http://aws.amazon.com/ec2/>

ATHENA WP3, McKenna, G. & Wyms, R. (2011). Persistent Identifiers (PIDs): Recommendations for Institutions. <http://www.athenaeurope.org/index.php?en/1/home>.

Bastijns, P., Coppens, S., Corneillie, S., Hochstenbach, P., Mannens, E., van Melle, A. (2009). BOM – Vlaanderen (Meta)datastandaarden voor digitale archieven. Universiteitsbibliotheek Gent. P.71-177.

Becker, C., Kulovits, H., Rauber, A., Hofman, H. (2008). Plato: A Service Oriented Decision Support System for Preservation Planning. Joint Conference on Digital Libraries '08, June 16-20, 2008, Pittsburgh, Pennsylvania, USA.

Becker, C., Kulovits, H., Guttenbrunner, M., Strodl, S., Rauber, A. & Hofman, H. (2009). Systematic planning for digital preservation: evaluating potential strategies and building preservation plans (Online). Springer Verlag. 19p. International Journal on Digital Libraries. Vol. 10, number 4. <http://www.ifs.tuwien.ac.at/~becker/pubs/becker-ijdl2009.pdf>. (February 2010).

Becker, C. & Rauber, A. (2011). Four cases, three solutions: Preservation plans for images. Vienna University of Technology, Austria (Online). <http://www.ifs.tuwien.ac.at/~becker/pubs/becker-four2011.pdf> (November 2011).

Bellini, E., Cirinna, C. & Lunghi, M. (2?) Persistent Identifiers for Cultural Heritage, briefing paper. Digital Preservation Europe, 4p (Online). http://www.digitalpreservationeurope.eu/publications/briefs/persistent_identifiers.pdf (June 2011).

Bradley, K. Editor (2009). Guidelines on the Production and Preservation of Digital Audio Objects, 2nd Edition. IASA – International Association of Sound and Audiovisual Archives.

Brown, A. (2008). Digital Preservation Guidance Note: 1-Selecting file formats for long-term preservation. National Archives – UK (Online). 10p. <http://www.nationalarchives.gov.uk/documents/selecting-file-formats.pdf> (April 2011).

Buyya, R., Yeo, C. S., Broberg, J. & Brandic, I. (2008). Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. Future Generation Computer Systems. Doi:10.1016. www.elsevier.com/locate/fgcs (April 2011).

Caplan, P. (2009). Understanding PREMIS. Library of Congress Network Development and MARC Standards Office (Online). <http://www.loc.gov/standards/premis/> (November 2011).

Caplan, P. (2011). PREMIS (LOC Preservation Metadata Maintenance Activity) OWL Ontology Available for Review (Online) <http://duraspace.org/premis-loc-preservation-metadata-maintenance-activity-owl-ontology-available-review> (November 2011).

CCSDS (2002). Recommendation for space data system standards Reference Model for an Open Archival Information System (OAIS) (Online). [Http://public.ccsds.org/publications/archive/650x0b1.pdf](http://public.ccsds.org/publications/archive/650x0b1.pdf) (June 2011).

Chen, F., Koufaty, D. A. & Zhang, X. (2009). Understanding Intrinsic Characteristics and System Implications of Flash Memory based Solid State Drives. In: SIGMETRICS/ performance '09, Proceedings of the eleventh international joint conference on Measurement and modelling of computer systems, Seattle, WA, USA.

Coalson, J. (2008). FLAC: Free Lossless Audio Codec (Online). <http://flac.sourceforge.net/documentation.html>. (April 2011).

Dappert, A. & Enders, M. (2010). Digital Preservation Metadata Standards. *Information Standards Quarterly*. Vol. 22, no. 2, 2010. National Institute Standards Organisation (NISO).

DCMI (2009). Guidelines for Dublin Core Application Profiles. <http://dublincore.org/documents/2009/05/18/profile-guidelines/> (September 2011).

Digitaal Erfgoed Nederland (Digital Heritage Netherlands), Goed Digitaliseren (Online). www.den.nl/digitaliseren (November 2011).

Dollar, C. M. & Ashley, L. (2009). Digital Preservation: A New Framework Based on a Capability Maturity Model. Presentation at the Managing Electronic Records Conference (Cohasset), May 19, 2009.

DPE, Digital Preservation Europe (2010). Preserving the British Library's C19 Newspaper Collection: Planets Film Released (Video). <http://www.digitalpreservationeurope.eu/> (October 2011).

Emmett, J. (2000). Five Years in the History of Audio Files. EBU Technical Review, December 2000 (Online). http://tech.ebu.ch/docs/techreview/trev_285-emmett.pdf (June 2011).

Ferreira, P. (2010). MXF – A technical overview. EBU technical review (Online). http://tech.ebu.ch/docs/techreview/trev_2010-Q3_MXF-2.pdf (May 2011).

Gerrard, R. (2004). Old Wine in New Skins: Migrating Legacy Data in a Museum Context. In: Preservation of Electronic Records: New Knowledge and Decision-making, Symposium CCI, Ottawa, Canada. P.135-141.

Gilbert, S. & Lynch, N. (2002). Brewer's Conjecture and the Feasibility of Consistent, Available, Partition-Tolerant Web Services. ACM SIGACT News, Volume 33, issue 2 (2002). P. 51-59.

Gladney, H. M & Lorie, R. (2005). Trustworthy 100-year digital objects: Durable encoding for when it's too late to ask. ACM Trans. Info. Sys. 23, 3. P. 299-324.

Gladney, H. M. (2006). Principles For Digital Preservation. Communications of the ACM, February 2006, Vol. 49, No2. P.111-116.

IBM (2011). IBMSMARTCLOUD (Online). <http://www.ibm.com/cloud-computing/us/en/> (April 2011)

ISO (1975). The set of control characters for ISO 646. Internet Assigned Numbers Authority Registry (Online). <http://www.itscj.ipsj.or.jp/ISO-IR/001.pdf> (May 2011).

ISO/IEC 15444-12 (2005). Information Technology – JPEG 2000 image coding system – Part 12: ISO base media file format. <http://www.jpeg.org/jpeg2000/index.html> (May 2011).

ISO/IEC 15444-3 (2007). Information technology - JPEG 2000 image coding system: Part 3: Motion JPEG 2000. <http://www.jpeg.org/jpeg2000/index.html> (May 2011).

ISO/IEC19005-1 (2005). Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF 1.4 (PDF/A-1). 36p. Document Format for Office Applications (Open Document) v1.0.

JISC Digital Media (2007). Using Optical Media for Digital Preservation (Online). <http://www.jiscdigitalmedia.ac.uk/stillimages/advice/using-optical-media-for-digital-preservation/> (December 2011).

Lawrence, G. W., Kehoe, W. R., Rieger, O. Y., Walters, W. H. & Kenney, A. R. (2000). Risk Management of Digital Information: A File Format Investigation. Washington D.C.: Council on Library and Information Resources (Online). <http://www.clir.org/pubs/reports/pub93/pub93.pdf> (September 2011).

Lee, K-H, Slattery, O., Lu, R., Tang, X. & McCrary, V. (2002). The State of the Art Practice in Digital Preservation. Journal of Research of the National Institute of Standards and Technology. Volume 107, issue 1 (2002). P.93–106.

Levy, S. (2011). A Cloudy Future. Wired Magazine, April 2011.

Library of Congress – National Digital Information Infrastructure & Preservation program (NDIIPP) (2007). AIFF (Audio Interchange File Format) (Online).
[Http://www.digitalpreservation.gov/formats/fdd/fdd000005.shtml#specs](http://www.digitalpreservation.gov/formats/fdd/fdd000005.shtml#specs) (September 2011).

Library of Congress. Formats (Online).
http://www.digitalpreservation.gov/formats/fdd/browse_list.shtml (November 2011).

Microsoft (2011). Power Cloud (Online). http://www.microsoft.com/en-us/cloud/default.aspx?fbid=iRE6B_-Datw (November 2011).

National Institute Standards Organisation (NISO) (2004). Understanding Metadata. NISO press, USA (Online). <http://www.niso.org/publications/press/UnderstandingMetadata.pdf> (November 2011).

National Library of New Zealand (NLNZ) (2003). Metadata Standards Framework – Preservation Metadata (Revised) (Online). <http://www.natlib.govt.nz/downloads/metaschema-revised.pdf> (September 2011).

OCLC/RLG Working Group on Preservation Metadata (2001). Preservation Metadata for Digital Objects: A Review of the State of the Art (Online).
http://www.oclc.org/research/activities/past/orprojects/pmwg/presmeta_wp.pdf (November 2011).

Plato (2010). Plato Preservation Planning Tool: Plato 3.0, User manual v.01 (Online)
http://www.ifs.tuwien.ac.at/dp/plato/docs/Plato_3_UserManual.pdf (2010).

PREMIS (2011). Data Dictionary for Preservation Metadata version 2.1(Online).
<http://www.loc.gov/standards/premis/> (November 2011).

RLG- National Archives and Records Administration Digital Repository Certification Task The Rockefeller Archive Center (2009). The Collaborative Electronic Records Project (Online).
<http://siarchives.si.edu/ceerp/tools.htm> (April 2011).

Rothenberg, J. (1999). Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation (Online). Publication 77. <http://db.grussell.org/section008.html> (April 2011).

Russell, G. (2008). Normalisation. Napier University, Edinburgh (Online).
<http://db.grussell.org/section008.html> (April 2011).

Schmalen, L. & Vary, P. (2009). Near-Lossless Compression and Protection by Turbo Source-Channel (De-) Coding using Irregular Index Assignments. Institute of Communication Systems and Data Processing, RWTH Aachen University, Germany (Online). <http://www.ind.rwth-aachen.de/fileadmin/publications/schmalen09.pdf> (November 2011).

Strodl, S., Becker, C., Neumayer, R. & Rauber, A. (2007). How to Choose a Digital Preservation Strategy: Evaluating a Preservation Planning Procedure, Vienna University of Technology, Austria (Online). www.ifs.tuwien.ac.at/dp (November 2011).

Task Force on Archiving of Digital Information (1996). Preserving Digital Information. Report to the Commission on Preservation and Access and the Research Libraries Group. Washington, D.C.

Testbed, Nationaal Archief of the Netherlands (2005). Cost of Digital Preservation. Digitale Bewaring, Testbed (Online). www.digitaleduurzaamheid.nl/bibliotheek/docs/CoDPv1.pdf (April 2011).

Verdegem, R. & van der Hoven, J. (2006). Emulation: To Be or Not To Be. In: Archiving 2006. Final Program and Proceedings, Ottawa, Canada. P. 56-60.

Verheul, I. (2006). Networking for digital preservation: Current practice in 15 libraries. IFLA Publications, Germany. P.45-67.

Wang, L. & Laszewski, G. von (2008). Scientific Cloud Computing: Early Definition and Experience. Service Oriented Cyberinfrastructure Lab, Rochester Institute of Technology (Online). <http://cyberaide.googlecode.com/svn/trunk/papers/08-cloud/vonLaszewski-08-cloud.pdf> (April 2011).

Weir, R. and Brauer, M. (2011). OASIS Open Document Format for Office Applications (OpenDocument) TC (Online). http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=office (November 2011).

ANNEX I - Digital Preservation Capability Performance Metrics

A sheet for filling in the model can be found below, after the 15 topics.

1. Policy

Value	Description
0	A formal written digital preservation policy does not exist.
1	A formal written digital preservation policy exists but it has not been officially issued.
2	A formal written digital preservation policy has been issued and widely disseminated.
3	A formal written digital preservation policy has been issued and widely disseminated. The policy is reviewed every two years for revision in light of experience with the policy.
4	A formal written digital preservation policy has been issued and widely disseminated. It is continuously reviewed for updates in light of changing circumstances so that it is considered a model for other digital preservation programmes.

2. Strategy

Value	Description
0	A systematic digital preservation strategy does not exist or if it does so, is not implemented.
1	A digital preservation strategy is in place that (1) keeps the bit stream of digital documents alive through planned device/media renewal and (2) retains digital records in their original format with the expectation that new software will be developed to support the original formats.
2	A digital preservation strategy is in place that (1) keeps the bit stream of digital documents alive through planned device/media renewal and (2) mitigates technology obsolescence through migration of digital records to technology neutral open standard formats.
3	A digital preservation strategy is in place that (1) keeps the bit stream of digital documents alive through planned device/media renewal, (2) mitigates technology obsolescence through normalisation of some digital records to technology neutral open standards formats, and (3) mitigates technology obsolescence through emulation of computer hardware.
4	A digital preservation strategy that (1) keeps the bit stream of digital documents alive through planned device/media renewal, (2) normalises digital records to technology neutral open standards formats, and (3) mitigates technology obsolescence through emulation of computer hardware and is continuously monitored for modifications and enhancements as technologies change. This digital preservation strategy serves as a model for other digital preservation programmes.

3. Governance Through Identified Roles and Responsibilities

Value	Description
0	No official mandate and supporting procedures for the long-term digital preservation of records that identify roles and responsibilities exist.
1	An official mandate and supporting procedures for the long-term

	preservation of records that identify roles and responsibilities exist but have not been implemented.
2	An official mandate and supporting procedures for the long-term preservation of records that identify roles and responsibilities has been implemented but the business unit that owns the digital preservation programme is not identified.
3	An official mandate and supporting procedures for the long-term preservation of records that identify roles and responsibilities has been implemented and the business unit that owns the digital preservation programme is identified.
4	An official mandate and supporting procedures for the long-term preservation of records that identify roles and responsibilities has been implemented and the business unit that owns the digital preservation programme is identified. The roles and responsibilities are periodically reviewed and revised in light of changing business needs and information technology capabilities.

4. Collaborative Engagement

Value	Description
0	There is little or no enterprise awareness of the importance of digital preservation for an organisation. If digital preservation projects exist they do not take into account the potential for collaboration with other parties which have an interest in digital preservation.
1	There is sufficient awareness of the role of collaboration in digital preservation that a collaborative framework is established within which some digital preservation projects may be undertaken.
2	There is sufficient awareness of the role of collaboration in digital preservation that a collaborative framework is established within which many digital preservation projects are undertaken.
3	A robust collaborative framework is in place that supports an enterprise digital preservation programme that is mission critical to the state within which digital preservation initiatives and activities are undertaken.
4	A robust collaborative framework is continuously monitored and updated to support an aggressive outreach to new collaboration partners that come into existence as technologies and organisations undergo change. It is considered a model for other digital preservation programmes.

5. Technical Expertise

Value	Description
0	Little or no professional expertise in digital preservation exists.
1	Sufficient professional technical expertise to support a minimal programme of digital preservation exists.
2	Sufficient professional technical expertise to support an intermediate programme of digital preservation exists.
3	Sufficient professional technical expertise to support an advanced programme of digital preservation exists.
4	Sufficient professional technical expertise to support an optimum programme of digital preservation exists. This level of professional technical exemplar is considered a model for other digital preservation programmes.

6. Open Source Software and Technology Neutral Open Standard Formats

Value	Description
0	No technology neutral open standard format that preserves the content and structure of digital records nor has the concept of digital preservation open source software been adopted.
1	One textual and one digital image technology neutral open standard format for digital preservation have been adopted to protect the content and structure of digital records. The concept of digital preservation open source software has been adopted but no digital preservation open source software is implemented.
2	A sufficient number of technology neutral open standard formats have been adopted to protect the content and structure of digital records that are likely to be transferred to the digital repository. The concept of digital preservation open source software has been adopted but no digital preservation open source software is implemented...
3	A sufficient number of technology neutral formats to protect the content and structure of digital records that are likely to be transferred to the digital repository have been adopted and is augmented by promoting the uses of these technology neutral open standard file formats at the time digital records are created. The concept of open source digital preservation software has been adopted and at least one instance of open source digital preservation software has been implemented.
4	Emerging technology neutral open standard formats are continuously monitored and adopted as appropriate. Entities that create and maintain digital records of long-term value are required to use designated technology neutral open standard formats and the digital repository employs only open source digital preservation software. This reliance on technology neutral open standard formats and open source software serves as a model for other digital preservation programmes.

7. Designated Communities

Value	Description
0	No written documentation exists that identifies target producers of records or defines their roles and obligations or target users of digital records.
1	Internal written documentation exists that identifies target producers and users of digital records and defines their roles and obligations at a very high level but with insufficient detail to support high level ingest and access processes.
2	Publicly available written documentation exists that identifies target producers and users of digital records and defines their roles and obligations in sufficient detail to support only a few detailed ingest and access processes.
3	Formal written agreements with producers of digital records that define their roles and obligations in considerable detail to support comprehensive ingest processes.
4	Formal written agreements with producers of digital records that define their roles and obligations in considerable detail to support comprehensive ingest processes. Publicly available written documentation defines the repository's access procedures that have been developed in conjunction with target user groups. These ingest and access processes are regularly reviewed and updated to take into

	account changing digital preservation best practices in this regard. These ingest and access processes serve as a model to other digital preservation programmes.
--	---

8. Digital Records Survey

Value	Description
0	No digital record survey protocol is in place to identify the scope and quantity of digital records that an organisation must preserve.
1	A digital record survey protocol is in place that is limited to an ad hoc analysis of a record schedule or anecdotal information to identify the scope and quantity of some (25%) of the digital records that an organisation must preserve.
2	A digital record survey protocol is in place that involves an analysis of a record schedule or other written documentation as well as information collected in interviews and questionnaires that identify the scope and quantity of many (50%) of the digital records that an organisation must preserve.
3	A digital record survey is in place that involves a comprehensive analysis of a record schedule or other written documentation as well as information collected through interviews and questionnaires that identify the scope and quantity of most (75%) of the digital records that an organisation must preserve. The analysis is updated every two years.
4	A digital record survey is conducted on a regular basis that involves an on-going comprehensive analysis of a record schedule or other written documentation as well as information collected in interviews and questionnaires that identify the scope and quantity of virtually all (95%) of the digital records that an organisation must preserve to satisfy the requirements of all stakeholders. The analysis is updated every year.

9. Compliance with ISO 14721 and TRACC Ingest Requirements

Value	Description
0	No tools or technologies are in place to conduct virus checks, validate and normalise file formats or render metadata that may be associated with the context of creation and use of digital records transferred to the repository
1	Repository tools and technologies are in place to conduct virus checks, validate and normalise file formats, and transfer digital records to new storage devices/media. Metadata is captured or created that establishes a narrow context of who created and used the digital records.
2	Repository tools and technologies are in place to conduct virus checks, validate and normalise file formats, transfer digital records to new storage devices/media, and normalise digital records containing text, selected structured data (spreadsheets), and digital images. Metadata is manually captured or created to establish a broad context of who created and used the digital records, and relationships with other digital records.
3	Repository tools and technologies are in place to conduct virus checks, validate and normalise file formats, transfer digital records to new storage devices/media, and normalised digital records containing text, selected structured data (spreadsheets), digital images, and vector drawings. Most metadata is automatically captured or created that establishes a broad context of who created and used the digital records, and relationships with other digital records.

4	Repository tools and technologies are in place to conduct virus checks, validate and normalise file formats, transfer digital records to new storage devices/media, and normalise digital records containing text, structured data (spreadsheets and databases), digital images, vector drawings, and web sites. All metadata is captured or created to establish a broad context of who created and used the digital records, and relationships with other digital records.
---	--

10. Storage Management

Value	Description
0	No dedicated logical or physical digital preservation repository exists.
1	A single copy of digital records is stored on desktop applications and removable storage media (e.g., CD or DVD).
2	Two copies of digital records are stored on logical or physical network storage devices that are maintained at two separate locations.
3	The storage of the two copies of digital records stored on logical or physical network storage repositories and maintained at two separate locations is supplemented with 'dark archives' at a third location. Synchronising digital preservation activities and documenting that all activities have been successfully completed ensure the referential integrity of all three copies.
4	The logical and physical storage of digital records at three different locations are continuously monitored and adjusted by implementing new storage practices and technologies as they emerge. This focus on storage management is a model for other digital preservation programmes.

11. Planned Device & Media Renewal

Value	Description
0	No formal device and media renewal procedure is in place.
1	A device and media renewal procedure is in place that calls for renewal of storage media when they are on the verge of becoming obsolescent.
2	A device and media renewal procedure is in place that requires renewal of storage media every ten years.
3	A device and media renewal procedure that renews device/storage media every ten years is supplemented by an annual media inspection programme to identify storage media that face imminent catastrophic loss.
4	A device and media renewal programme is in place that continuously monitors the potential loss of digital record readability and automatically writes them to new storage media as necessary. The media renewal programme is a model for other digital preservation programs.

12. Digital Records Integrity

Level	Description
0	No procedure is in place to validate the integrity of digital records.
1	Comparison of bit/byte counts before and after preservation activities are used to validate the integrity of digital records.
2	Comparison of MD5 hash digests before and after preservation activities to validate the integrity of digital records.
3	Comparison of SHA-2 hash digests before and after preservation activities to validate the integrity of digital records.
4	Digital records are encapsulated and digitally signed after each preservation

	action to validate the integrity. The integrity validation process in use is continuously evaluated and updated as new processes available. The techniques used to validate the integrity of digital records serve as a model to other digital preservation programmes.
--	---

13. Digital Records Security

Value	Description
0	No formal disaster recovery, business resumption, backups, and physical security processes are in place to protect digital records.
1	Disaster recovery, backups, and physical security processes that protect the security of digital records are maintained through desktop applications and removable storage media (e.g., CD or DVD)
2	Disaster recovery, business resumption and backup processes are network based. Physical security is assured through role based permission access.
3	Digital records are written to non-rewritable storage media that are protected by network based disaster recovery, business resumption and backups. Physical security is assured through role based permission access.
4	Digital record security processes are continuously monitored and innovations are introduced as appropriate. These digital record security approaches and techniques serve as a model to other digital preservation programmes.

14. Digital Preservation Metadata

Value	Description
0	Little or no preservation metadata associated with records of long-term value is collected and maintained.
1	Preservation metadata for digital records of long-term value is collected and maintained on an ad hoc basis.
2	Preservation metadata for some digital records of long-term value is collected on a systematic basis in accordance with established guidelines and protected at the same level as the digital records associated with the metadata.
3	Preservation metadata for most digital records of long-term value is collected on a systematic basis in accordance with established guidelines and protected at the same level as the digital records associated with the metadata.
4	Preservation metadata for all digital records of long-term value is collected on a systematic basis in accordance with established guidelines and protected at the same level as the digital records associated with the metadata. Preservation metadata collection guidelines are continuously reviewed and updated as required. These metadata preservation processes and guidelines serve as a model to other digital preservation programmes.

15. Access to Digital Records

Value	Description
0	There is little or no electronic access to digital records of long-term value. In some instances hard copies of digital records may be requested electronically.

1	A few digital records of long-term value are electronically accessible but are only available in ASCII text.
2	Some digital records of long-term value are electronically accessible but only in ASCII text, TIFF images, or PDF.
3	Most digital records of long-term value are electronically accessible in technology neutral open standard formats for text, images, and vector graphics.
4	All digital records of long-term value are electronically accessible in any format the user community requires. The tools used to support electronic accessibility are continuously reviewed and updated to reflect technology changes or the needs of the user community. Electronic accessibility processes and guidelines serve as a model to other digital preservation programmes.

Digital Preservation Readiness Capability Maturity Model Assessment

A major objective of the Digital Preservation Readiness Capability Maturity Model is to identify the current state of digital preservation readiness capability of an organisation at a high level. To accomplish this objective it is necessary to collect and analyse information about the state of digital preservation of an organisation and map this information into a matrix that is organised by the 15 performance metrics and the five stages of the Digital Preservation Readiness Capability Maturity Model. Aggregating these numerical scores yields a composite digital preservation readiness numerical score that falls into one of the five categories listed below.

- Nominal 0 - 15
- Minimal 16 - 30
- Intermediate 31 - 45
- Advanced 46 - 60
- Optimum 61 - 75

Topic	0	1	2	3	4
1. Designated Communities					
2. Collaborative Engagement					
3. Governance Through Identified Roles and Responsibilities					
4. Policy					
5. Strategy					
6. Digital Record Survey					
7. Storage Management					
8. Digital Record Ingest					
9. Digital Records Security					
10. Planned Device & Media Renewal					
11. Technical Expertise					
12. Access to Digital Records					
13. Digital Preservation Metadata					
14. Digital Record Integrity					
15. Open Sources and Technology Neutral Open Standard Formats					

ANNEX II - File formats and codec suitable for Web and presentation

Audio		
Ogg Vorbis	Ogg Vorbis is an open source, platform independent, lossy compression format for sound files. It was developed as an alternative to the MP3.	Presentation and Web
mp3	In the MPEG family, but specific audio format. It is a standardised, lossy compression format, and is also very well known and widespread.	Presentation and Web
Text		
XML	Extensible Mark-up Language is a subset from SGML, making the text suitable for the Internet. Using XML, the structure of text documents are defined using tags and attributes. XML can also be used to describe data that would normally be included in relational database systems. (www.den.nl)	Web
ePUB	Electronic Publication is a XML based standard for digital books and articles. The text is scalable, adapting the size to the screen on which it is shown on (www.den.nl).	Presentation
UTF-8 ¹	Backwards compatible with ASCII. It is the dominant character encoder for the Web.	Web
HTML5	HTML5 is a platform independent Web document codec, it can include audio-visual materials as well as texts and images.	Web
Video and video containers		
MOV	Multimedia format, made by Apple, originally for Quicktime frameworks. It is a proprietary container but is widely used for video and audio content.	Presentation
V210	AJA Video Systems have implemented this Quicktime video codec for Windows. It is a 10 bit per component, YCrCb 4:2:2 format in which samples for 5 pixels are packed into 4 4-byte little-endian words.	Presentation
Ogg Theora	Ogg Theoram is an open source, lossy compressed video format made for the internet, but is not supported by Internet Explorer.	Web

FLV	Flash Video is a container format for audio-visual material on the Web. The format is open, but the codec is patented, which makes it not completely open source. Some platforms are not compatible with Flash Video, such as iPads and iPhones. (www.den.nl)	Web
MPEG	Moving Picture Experts Group has developed several standards for mainly lossy compressed, audio-visual digital material. MPEG files are open source, but often still require licenses. The MPEGs can be used for presentation purposes, but are not recommended as preservation formats.	Open standard, widespread, lossless and lossy content, platform independent
MPEG4 ²	ISO standard for multimedia applications introduced in 1998. MPEG-4 was created to stream DVD quality video at lower data rates and with smaller file sizes. MPEG-4 is a well-known and often used format. Although it is made a standard by ISO, it still requires a license (www.den.nl/standaard/194/Moving-Pictures-experts-Group-Mpeg-4).	Presentation
MPEG7 ³⁴	MPEG-7 offers a comprehensive set of audio-visual metadata description tools, and is not in itself an audio-visual format. The MPEG group calls it a content representation standard for multimedia information search and filtering.	Metadata format

1 ISO/IEC Information technology -- Universal Multiple-Octet Coded Character Set (UCS) -- Part 1: Architecture and Basic Multilingual Plane

2 Koenen, R. (2002). Overview of the MPEG-4 standard. ISO/IEC JTC1/SC29/WG11 N4668.

<http://mpeg.chiariglione.org/standards/mpeg-4/mpeg-4.htm>

3 Martinez, J. M. (2004). MPEG-7 Overview. ISO/IEC JTC1/SC29/WG11N6828.

<http://mpeg.chiariglione.org/standards/mpeg-7/mpeg-7.htm#E9E3>

4 Martinez, J. M., Koenen, R. & Pereira, F. (2002). Standards. MPEG-7, The Generic Multimedia Content Description Standard, part1. 1070-986X/02/\$17.00 © 2002 IEEE (Online). <http://www.coolutils.com/formats/mpeg>

ANNEX III - Manual for metadata extraction

The software needed can be found by DCA partners on the Mybbt communication platform and downloaded from there. One can easily find ExifTools on the Web (<http://www.sno.phy.queensu.ca/~phil/exiftool/>).

Metadata extraction for Windows

- Download ExifTool for Windows
- Open 'exiftool (windows).zip'.
- Copy 'exiftool(-k -a -u -g1 -w txt).exe' to a convenient place, the desktop for instance.
- IMPORTANT: do *NOT* rename the file; the way the file is named defines its behaviour.
- Now one can drop any video, audio or image file onto the ExifTool icon (the camel) and a text file with all the technical metadata will appear next to the media file.
- For example, if one drops birds.jpg onto the ExifTool icon, a file birds.txt will be created in the directory of birds.jpg.

Metadata extraction for Mac

- Download and install 'ExifTool (mac)-8.68.dmg'.
 - Download and unzip 'exiftool (mac automator).zip'.
 - Copy or move 'exiftool (mac automator)' to a convenient place, the desktop for instance.
- Now one can drop any video, audio or image file onto the ExifTool icon and a text file with all the technical metadata will appear next to the media file.
- For example, if one drops birds.jpg onto the ExifTool icon, a file birds.txt will be created in the directory of birds.jpg.

ANNEX IV - Example of a metadata schema with PREMIS

(Source: <http://www.loc.gov/standards/premis/louis-2-0.xml>)

This XML file does not appear to have any style information associated to it. The document tree is shown below.

```
<mets:mets xmlns:lc="http://www.loc.gov/mets/profiles/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance" xmlns:mods="http://www.loc.gov/mods/v3" xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:mets
="http://www.loc.gov/METS/" xmlns:photo="http://www.loc.gov/mets/profiles/photoObject" xmlns:premis="inf
o:lc/xmlns/premis-
v2" PROFILE="lc:photoObject" OBJID="loc.natlib.gottlieb.09601" xsi:schemaLocation="http://www.loc.gov/
METS/ http://www.loc.gov/standards/mets/mets.xsd">
<mets:dmdSec ID="MODS">
<mets:mdWrap MDTYPE="MODS">
<mets:xmlData>
<mods:mods xmlns:mods="http://www.loc.gov/mods/v3" xsi:schemaLocation="http://www.loc.gov/mods/v3
http://www.loc.gov/standards/mods/v3/mods-3-3.xsd" ID="ver1">
<mods:titleInfo>
<mods:title>
[Portrait of Louis Armstrong, between 1938 and 1948]
</mods:title>
</mods:titleInfo>
<mods:name type="personal">
<mods:namePart>Gottlieb, William P.</mods:namePart>
<mods:namePart type="date">1917-</mods:namePart>
<mods:role>
<mods:roleTerm authority="marcrelator" type="text">creator</mods:roleTerm>
</mods:role>
<mods:role>
<mods:roleTerm type="text">photographer.</mods:roleTerm>
</mods:role>
</mods:name>
<mods:typeOfResource>still image</mods:typeOfResource>
<mods:genre authority="marc">photograph</mods:genre>
<mods:genre authority="gmgpc">Portrait photographs-1930-1950.</mods:genre>
<mods:genre authority="gmgpc">Film negatives-1930-1950.</mods:genre>
<mods:originInfo>
<mods:place>
<mods:placeTerm type="code" authority="marccountry">xxu</mods:placeTerm>
</mods:place>
<mods:dateIssued encoding="marc" point="start">1938</mods:dateIssued>
<mods:dateIssued encoding="marc" point="end">1948</mods:dateIssued>
<mods:dateIssued encoding="marc" point="start" qualifier="questionable">1938</mods:dateIssued>
<mods:dateIssued encoding="marc" point="end" qualifier="questionable">1948</mods:dateIssued>
<mods:issuance>monographic</mods:issuance>
</mods:originInfo>
<mods:physicalDescription>
<mods:form authority="gmd">graphic</mods:form>
<mods:extent>1 negative : b&w ; 3 1/4 x 4 1/4 in.</mods:extent>
</mods:physicalDescription>
<mods:note>Gottlieb Collection Assignment No. 040</mods:note>
<mods:note>Original negative and contact print not served.</mods:note>
<mods:note>Purchase William P. Gottlieb</mods:note>
<mods:note type="version">original negative</mods:note>
<mods:subject authority="lcsch">
<mods:name type="personal">
<mods:namePart>Armstrong, Louis, 1900-1971</mods:namePart>
</mods:name>
</mods:subject>
```

```

<mods:subject authority="lcs" >
<mods:topic>Jazz musicians</mods:topic>
<mods:temporal>1930-1950</mods:temporal>
</mods:subject>
<mods:subject authority="lcs" >
<mods:topic>Trumpet players</mods:topic>
<mods:temporal>1930-1950</mods:temporal>
</mods:subject>
<mods:classification authority="lcc">LC-GLB13- 0960</mods:classification>
<mods:location>
<mods:physicalLocation displayLabel="negative">
Library of Congress Prints & Photographs Division Washington D.C. 20540 USA
</mods:physicalLocation>
</mods:location>
<mods:location>
<mods:physicalLocation displayLabel="contact print">
Library of Congress Prints & Photographs Division Washington D.C. 20540 USA
</mods:physicalLocation>
</mods:location>
<mods:identifier type="stock number">LC-GLB13-0960 DLC</mods:identifier>
<mods:accessCondition type="restrictionOnAccess">Original negative and contact print not
served.</mods:accessCondition>
<mods:recordInfo>
<mods:recordContentSource authority="marcorg">DLC</mods:recordContentSource>
<mods:recordCreationDate encoding="marc">990119</mods:recordCreationDate>
<mods:recordChangeDate encoding="iso8601">19990520104721.0</mods:recordChangeDate>
<mods:recordIdentifier source="DLC">got99000960</mods:recordIdentifier>
</mods:recordInfo>
<mods:relatedItem type="otherVersion" ID="ver2">
<mods:note type="version">contact print with annotations</mods:note>
</mods:relatedItem>
</mods:mods>
</mets:xmlData>
</mets:mdWrap>
</mets:dmdSec>
<mets:amdSec>
<mets:techMD ID="object1">
<mets:mdWrap MDTYPE="PREMIS:OBJECT">
<mets:xmlData>
<premis:object xsi:type="premis:file" xsi:schemaLocation="info:lc/xmlns/premis-v2
http://www.loc.gov/standards/premis/v2/premis-v2-0.xsd">
<premis:objectIdentifier>
<premis:objectIdentifierType>hdl</premis:objectIdentifierType>
<premis:objectIdentifierValue>loc.music/gottlieb.09601</premis:objectIdentifierValue>
</premis:objectIdentifier>
<premis:preservationLevel>
<premis:preservationLevelValue>full</premis:preservationLevelValue>
<premis:preservationLevelDateAssigned>20070529</premis:preservationLevelDateAssigned>
</premis:preservationLevel>
<premis:significantProperties>
<premis:significantPropertiesType>behavior</premis:significantPropertiesType>
<premis:significantPropertiesValue>hyperlinks traversable</premis:significantPropertiesValue>
</premis:significantProperties>
<premis:objectCharacteristics>
<premis:compositionLevel>0</premis:compositionLevel>
<premis:fixity>
<premis:messageDigestAlgorithm>MD5</premis:messageDigestAlgorithm>
<premis:messageDigest>36b03197ad066cd719906c55eb68ab8d</premis:messageDigest>
<premis:messageDigestOriginator>LocalDCMS</premis:messageDigestOriginator>
</premis:fixity>

```

```

<premis:size>20800896</premis:size>
<premis:format>
<premis:formatDesignation>
<premis:formatName>image/tiff</premis:formatName>
<premis:formatVersion>6.0</premis:formatVersion>
</premis:formatDesignation>
<premis:formatRegistry>
<premis:formatRegistryName>PRONOM</premis:formatRegistryName>
<premis:formatRegistryKey>fmt/10</premis:formatRegistryKey>
<premis:formatRegistryRole>specification</premis:formatRegistryRole>
</premis:formatRegistry>
</premis:format>
<premis:creatingApplication>
<premis:creatingApplicationName>ScandAll 21</premis:creatingApplicationName>
<premis:creatingApplicationVersion>4.1.4</premis:creatingApplicationVersion>
<premis:dateCreatedByApplication>1998-10-30</premis:dateCreatedByApplication>
</premis:creatingApplication>
<premis:creatingApplication>
<premis:creatingApplicationName>Adobe Photoshop</premis:creatingApplicationName>
<premis:creatingApplicationVersion>CS2</premis:creatingApplicationVersion>
<premis:dateCreatedByApplication>2006-09-20T08:29:02</premis:dateCreatedByApplication>
</premis:creatingApplication>
<premis:objectCharacteristicsExtension>
<mix:mix xmlns:mix="http://www.loc.gov/mix/v20" xsi:schemaLocation="http://www.loc.gov/mix/v20
http://www.loc.gov/standards/mix/mix20/mix20.xsd">
<mix:BasicDigitalObjectInformation>
<mix:byteOrder>little endian</mix:byteOrder>
<mix:Compression>
<mix:compressionScheme>1</mix:compressionScheme>
</mix:Compression>
</mix:BasicDigitalObjectInformation>
<mix:BasicImageInformation>
<mix:BasicImageCharacteristics>
<mix:imageWidth>3982</mix:imageWidth>
<mix:imageHeight>5223</mix:imageHeight>
<mix:PhotometricInterpretation>
<mix:colorSpace>1</mix:colorSpace>
</mix:PhotometricInterpretation>
</mix:BasicImageCharacteristics>
</mix:BasicImageInformation>
<mix:ImageCaptureMetadata>
<mix:GeneralCaptureInformation>
<mix:dateTimeCreated>1998-10-03T08:25:28</mix:dateTimeCreated>
<mix:imageProducer>Library of Congress</mix:imageProducer>
</mix:GeneralCaptureInformation>
<mix:orientation>normal*</mix:orientation>
</mix:ImageCaptureMetadata>
<mix:ImageAssessmentMetadata>
<mix:SpatialMetrics>
<mix:samplingFrequencyUnit>no absolute unit of measurement</mix:samplingFrequencyUnit>
<mix:xSamplingFrequency>
<mix:numerator>3982</mix:numerator>
<mix:denominator>1</mix:denominator>
</mix:xSamplingFrequency>
<mix:ySamplingFrequency>
<mix:numerator>5223</mix:numerator>
<mix:denominator>1</mix:denominator>
</mix:ySamplingFrequency>
</mix:SpatialMetrics>
<mix:ImageColorEncoding>

```

```

<mix:BitsPerSample>
<mix:bitsPerSampleValue>8</mix:bitsPerSampleValue>
</mix:BitsPerSample>
<mix:samplesPerPixel>1</mix:samplesPerPixel>
</mix:ImageColorEncoding>
</mix:ImageAssessmentMetadata>
</mix:mix>
</premis:objectCharacteristicsExtension>
</premis:objectCharacteristics>
<premis:originalName>0001h.tif</premis:originalName>
<premis:storage>
<premis:contentLocation>
<premis:contentLocationType>filepath</premis:contentLocationType>
<premis:contentLocationValue>amserver</premis:contentLocationValue>
</premis:contentLocation>
<premis:storageMedium>disk</premis:storageMedium>
</premis:storage>
<premis:environment>
<premis:environmentCharacteristic>recommended</premis:environmentCharacteristic>
<premis:environmentPurpose>render</premis:environmentPurpose>
<premis:environmentPurpose>edit</premis:environmentPurpose>
<premis:software>
<premis:swName>Adobe Acrobat</premis:swName>
<premis:swVersion>5.0</premis:swVersion>
<premis:swType>renderer</premis:swType>
</premis:software>
<premis:software>
<premis:swName>Windows</premis:swName>
<premis:swVersion>XP</premis:swVersion>
<premis:swType>operatingSystem</premis:swType>
</premis:software>
<premis:hardware>
<premis:hwName>Intel x86</premis:hwName>
<premis:hwType>processor</premis:hwType>
<premis:hwOtherInformation>60 mhz minimum</premis:hwOtherInformation>
</premis:hardware>
<premis:hardware>
<premis:hwName>64 MB RAM</premis:hwName>
<premis:hwType>memory</premis:hwType>
<premis:hwOtherInformation>32 MB minimum</premis:hwOtherInformation>
</premis:hardware>
<premis:environmentExtension>
<hardwareInformation/>
<softwareInformation/>
</premis:environmentExtension>
</premis:environment>
<premis:relationship>
<premis:relationshipType>structural</premis:relationshipType>
<premis:relationshipSubType>is sibling</premis:relationshipSubType>
<premis:relatedObjectIdentification>
<premis:relatedObjectIdentifierType>hdl</premis:relatedObjectIdentifierType>
<premis:relatedObjectIdentifierValue>loc.music/gottlieb.09602</premis:relatedObjectIdentifierValue>
<premis:relatedObjectSequence>0</premis:relatedObjectSequence>
</premis:relatedObjectIdentification>
</premis:relationship>
<premis:relationship>
<premis:relationshipType>structural</premis:relationshipType>
<premis:relationshipSubType>is sibling</premis:relationshipSubType>
<premis:relatedObjectIdentification>
<premis:relatedObjectIdentifierType>URI</premis:relatedObjectIdentifierType>

```

```

<premis:relatedObjectIdentifierValue>
http://lcweb2.loc.gov/cocoon/ihhas/loc.natlib.gottlieb.09601/mets.xml
</premis:relatedObjectIdentifierValue>
<premis:relatedObjectSequence>0</premis:relatedObjectSequence>
</premis:relatedObjectIdentification>
</premis:relationship>
<premis:relationship>
<premis:relationshipType>derivation</premis:relationshipType>
<premis:relationshipSubType>is source of</premis:relationshipSubType>
<premis:relatedObjectIdentification>
<premis:relatedObjectIdentifierType>URL</premis:relatedObjectIdentifierType>
<premis:relatedObjectIdentifierValue>
http://lcweb2.loc.gov/natlib/ihhas/service/gottlieb/09601/ver01/0001v.jpg
</premis:relatedObjectIdentifierValue>
<premis:relatedObjectSequence>0</premis:relatedObjectSequence>
</premis:relatedObjectIdentification>
<premis:relatedEventIdentification>
<premis:relatedEventIdentifierType>LocalDCMS</premis:relatedEventIdentifierType>
<premis:relatedEventIdentifierValue>E002.1</premis:relatedEventIdentifierValue>
<premis:relatedEventSequence>1</premis:relatedEventSequence>
</premis:relatedEventIdentification>
</premis:relationship>
<premis:linkingEventIdentifier>
<premis:linkingEventIdentifierType>Local Repository</premis:linkingEventIdentifierType>
<premis:linkingEventIdentifierValue>E001.1</premis:linkingEventIdentifierValue>
</premis:linkingEventIdentifier>
<premis:linkingEventIdentifier>
<premis:linkingEventIdentifierType>Local Repository</premis:linkingEventIdentifierType>
<premis:linkingEventIdentifierValue>E001.2</premis:linkingEventIdentifierValue>
</premis:linkingEventIdentifier>
<premis:linkingIntellectualEntityIdentifier>
<premis:linkingIntellectualEntityIdentifierType>hdl</premis:linkingIntellectualEntityIdentifierType>
<premis:linkingIntellectualEntityIdentifierValue>loc.natlib.gottlieb.09601</premis:linkingIntellectualEntityIdentifierValue>
</premis:linkingIntellectualEntityIdentifier>
<premis:linkingIntellectualEntityIdentifier>
<premis:linkingIntellectualEntityIdentifierType>URI</premis:linkingIntellectualEntityIdentifierType>
<premis:linkingIntellectualEntityIdentifierValue>
http://lcweb2.loc.gov/cocoon/ihhas/loc.natlib.gottlieb.09601/default.html
</premis:linkingIntellectualEntityIdentifierValue>
</premis:linkingIntellectualEntityIdentifier>
</premis:object>
</mets:xmlData>
</mets:mdWrap>
</mets:techMD>
<mets:techMD ID="object2">
<mets:mdWrap MDTYPE="PREMIS:OBJECT">
<mets:xmlData>
<premis:object xsi:type="premis:file">
<premis:objectIdentifier>
<premis:objectIdentifierType>hdl</premis:objectIdentifierType>
<premis:objectIdentifierValue>loc.music/gottlieb.09602</premis:objectIdentifierValue>
</premis:objectIdentifier>
<premis:preservationLevel>
<premis:preservationLevelValue>full</premis:preservationLevelValue>
<premis:preservationLevelDateAssigned>20070529</premis:preservationLevelDateAssigned>
</premis:preservationLevel>
<premis:significantProperties>
<premis:significantPropertiesType>behavior</premis:significantPropertiesType>
<premis:significantPropertiesValue>hyperlinks traversable</premis:significantPropertiesValue>

```

```

</premis:significantProperties>
<premis:objectCharacteristics>
<premis:compositionLevel>0</premis:compositionLevel>
<premis:fixity>
<premis:messageDigestAlgorithm>MD5</premis:messageDigestAlgorithm>
<premis:messageDigest>ceb3dbc5dacd3883d0985174ef5df7db</premis:messageDigest>
<premis:messageDigestOriginator>LocalDCMS</premis:messageDigestOriginator>
</premis:fixity>
<premis:size>58238300</premis:size>
<premis:format>
<premis:formatDesignation>
<premis:formatName>image/tiff</premis:formatName>
<premis:formatVersion>6.0</premis:formatVersion>
</premis:formatDesignation>
<premis:formatRegistry>
<premis:formatRegistryName>PRONOM</premis:formatRegistryName>
<premis:formatRegistryKey>fmt/10</premis:formatRegistryKey>
<premis:formatRegistryRole>specification</premis:formatRegistryRole>
</premis:formatRegistry>
</premis:format>
<premis:creatingApplication>
<premis:creatingApplicationName>ScandAll 21</premis:creatingApplicationName>
<premis:creatingApplicationVersion>4.1.4</premis:creatingApplicationVersion>
<premis:dateCreatedByApplication>1998-10-30</premis:dateCreatedByApplication>
</premis:creatingApplication>
<premis:creatingApplication>
<premis:creatingApplicationName>Adobe Photoshop</premis:creatingApplicationName>
<premis:creatingApplicationVersion>CS2</premis:creatingApplicationVersion>
<premis:dateCreatedByApplication>2006-09-20T08:29:02</premis:dateCreatedByApplication>
</premis:creatingApplication>
<premis:objectCharacteristicsExtension>
<mix:mix xmlns:mix="http://www.loc.gov/mix/v20" xsi:schemaLocation="http://www.loc.gov/mix/v20
http://www.loc.gov/standards/mix/mix20/mix20.xsd">
<mix:BasicDigitalObjectInformation>
<mix:byteOrder>little endian</mix:byteOrder>
<mix:Compression>
<mix:compressionScheme>1</mix:compressionScheme>
</mix:Compression>
</mix:BasicDigitalObjectInformation>
<mix:BasicImageInformation>
<mix:BasicImageCharacteristics>
<mix:imageWidth>3982</mix:imageWidth>
<mix:imageHeight>5223</mix:imageHeight>
<mix:PhotometricInterpretation>
<mix:colorSpace>2</mix:colorSpace>
<mix:ReferenceBlackWhite>
<mix:Component>
<mix:componentPhotometricInterpretation>R</mix:componentPhotometricInterpretation>
<mix:footroom>
<mix:numerator>255</mix:numerator>
<mix:denominator>1</mix:denominator>
</mix:footroom>
<mix:headroom>
<mix:numerator>0</mix:numerator>
<mix:denominator>1</mix:denominator>
</mix:headroom>
</mix:Component>
<mix:Component>
<mix:componentPhotometricInterpretation>G</mix:componentPhotometricInterpretation>
<mix:footroom>

```

```

<mix:numerator>255</mix:numerator>
<mix:denominator>1</mix:denominator>
</mix:footroom>
<mix:headroom>
<mix:numerator>0</mix:numerator>
<mix:denominator>1</mix:denominator>
</mix:headroom>
</mix:Component>
<mix:Component>
<mix:componentPhotometricInterpretation>B</mix:componentPhotometricInterpretation>
<mix:footroom>
<mix:numerator>255</mix:numerator>
<mix:denominator>1</mix:denominator>
</mix:footroom>
<mix:headroom>
<mix:numerator>0</mix:numerator>
<mix:denominator>1</mix:denominator>
</mix:headroom>
</mix:Component>
</mix:ReferenceBlackWhite>
</mix:PhotometricInterpretation>
</mix:BasicImageCharacteristics>
</mix:BasicImageInformation>
<mix:ImageCaptureMetadata>
<mix:GeneralCaptureInformation>
<mix:dateTimeCreated>1998-10-30T08:29:02</mix:dateTimeCreated>
<mix:imageProducer>Library of Congress</mix:imageProducer>
</mix:GeneralCaptureInformation>
<mix:orientation>normal*</mix:orientation>
</mix:ImageCaptureMetadata>
<mix:ImageAssessmentMetadata>
<mix:SpatialMetrics>
<mix:samplingFrequencyUnit>no absolute unit of measurement</mix:samplingFrequencyUnit>
<mix:xSamplingFrequency>
<mix:numerator>3882</mix:numerator>
<mix:denominator>1</mix:denominator>
</mix:xSamplingFrequency>
<mix:ySamplingFrequency>
<mix:numerator>5000</mix:numerator>
<mix:denominator>1</mix:denominator>
</mix:ySamplingFrequency>
</mix:SpatialMetrics>
<mix:ImageColorEncoding>
<mix:BitsPerSample>
<mix:bitsPerSampleValue>8</mix:bitsPerSampleValue>
</mix:BitsPerSample>
<mix:samplesPerPixel>1</mix:samplesPerPixel>
</mix:ImageColorEncoding>
</mix:ImageAssessmentMetadata>
</mix:mix>
</premis:objectCharacteristicsExtension>
</premis:objectCharacteristics>
<premis:originalName>0002h.tif</premis:originalName>
<premis:storage>
<premis:contentLocation>
<premis:contentLocationType>filepath</premis:contentLocationType>
<premis:contentLocationValue>amserver/</premis:contentLocationValue>
</premis:contentLocation>
<premis:storageMedium>disk</premis:storageMedium>
</premis:storage>

```

```

<premis:environment>
<premis:environmentCharacteristic>recommended</premis:environmentCharacteristic>
<premis:environmentPurpose>render</premis:environmentPurpose>
<premis:environmentPurpose>edit</premis:environmentPurpose>
<premis:software>
<premis:swName>Adobe Acrobat</premis:swName>
<premis:swVersion>5.0</premis:swVersion>
<premis:swType>renderer</premis:swType>
</premis:software>
<premis:software>
<premis:swName>Windows</premis:swName>
<premis:swVersion>XP</premis:swVersion>
<premis:swType>operatingSystem</premis:swType>
</premis:software>
<premis:hardware>
<premis:hwName>Intel x86</premis:hwName>
<premis:hwType>processor</premis:hwType>
<premis:hwOtherInformation>60 mhz minimum</premis:hwOtherInformation>
</premis:hardware>
<premis:hardware>
<premis:hwName>64 MB RAM</premis:hwName>
<premis:hwType>memory</premis:hwType>
<premis:hwOtherInformation>32 MB minimum</premis:hwOtherInformation>
</premis:hardware>
<premis:environmentExtension>
<hardwareInformation/>
<softwareInformation/>
</premis:environmentExtension>
</premis:environment>
<premis:relationship>
<premis:relationshipType>structural</premis:relationshipType>
<premis:relationshipSubType>is sibling</premis:relationshipSubType>
<premis:relatedObjectIdentification>
<premis:relatedObjectIdentifierType>hdl</premis:relatedObjectIdentifierType>
<premis:relatedObjectIdentifierValue>loc.music/gottlieb.09601</premis:relatedObjectIdentifierValue>
<premis:relatedObjectSequence>0</premis:relatedObjectSequence>
</premis:relatedObjectIdentification>
</premis:relationship>
<premis:relationship>
<premis:relationshipType>structural</premis:relationshipType>
<premis:relationshipSubType>is sibling</premis:relationshipSubType>
<premis:relatedObjectIdentification>
<premis:relatedObjectIdentifierType>URI</premis:relatedObjectIdentifierType>
<premis:relatedObjectIdentifierValue>
http://lcweb2.loc.gov/cocoon/ihascocoon/loc.natlib.gottlieb.09601/mets.xml
</premis:relatedObjectIdentifierValue>
<premis:relatedObjectSequence>0</premis:relatedObjectSequence>
</premis:relatedObjectIdentification>
</premis:relationship>
<premis:relationship>
<premis:relationshipType>derivation</premis:relationshipType>
<premis:relationshipSubType>is source of</premis:relationshipSubType>
<premis:relatedObjectIdentification>
<premis:relatedObjectIdentifierType>URL</premis:relatedObjectIdentifierType>
<premis:relatedObjectIdentifierValue>
http://lcweb2.loc.gov/natl/lib/ihascocoon/service/gottlieb/09601/ver02/0001v.jpg
</premis:relatedObjectIdentifierValue>
<premis:relatedObjectSequence>0</premis:relatedObjectSequence>
</premis:relatedObjectIdentification>
<premis:relatedEventIdentification>

```

```

<premis:relatedEventIdentifierType>LocalDCMS</premis:relatedEventIdentifierType>
<premis:relatedEventIdentifierValue>E002.2</premis:relatedEventIdentifierValue>
<premis:relatedEventSequence>1</premis:relatedEventSequence>
</premis:relatedEventIdentification>
</premis:relationship>
<premis:linkingEventIdentifier>
<premis:linkingEventIdentifierType>Local Repository</premis:linkingEventIdentifierType>
<premis:linkingEventIdentifierValue>E001.3</premis:linkingEventIdentifierValue>
</premis:linkingEventIdentifier>
<premis:linkingEventIdentifier>
<premis:linkingEventIdentifierType>Local Repository</premis:linkingEventIdentifierType>
<premis:linkingEventIdentifierValue>E001.4</premis:linkingEventIdentifierValue>
</premis:linkingEventIdentifier>
<premis:linkingIntellectualEntityIdentifier>
<premis:linkingIntellectualEntityIdentifierType>hdl</premis:linkingIntellectualEntityIdentifierType>
<premis:linkingIntellectualEntityIdentifierValue>loc.natlib.gottlieb.09601</premis:linkingIntellectualEntityIdentifierValue>
</premis:linkingIntellectualEntityIdentifier>
<premis:linkingIntellectualEntityIdentifier>
<premis:linkingIntellectualEntityIdentifierType>URI</premis:linkingIntellectualEntityIdentifierType>
<premis:linkingIntellectualEntityIdentifierValue>
http://lcweb2.loc.gov/cocoon/ahas/loc.natlib.gottlieb.09601/default.html
</premis:linkingIntellectualEntityIdentifierValue>
</premis:linkingIntellectualEntityIdentifier>
</premis:object>
</mets:xmlData>
</mets:mdWrap>
</mets:techMD>
<mets:digiprovMD ID="event1">
<mets:mdWrap MDTYPE="PREMIS:EVENT">
<mets:xmlData>
<premis:event>
<premis:eventIdentifier>
<premis:eventIdentifierType>LocalRepository</premis:eventIdentifierType>
<premis:eventIdentifierValue>e001.1</premis:eventIdentifierValue>
</premis:eventIdentifier>
<premis:eventType>validation</premis:eventType>
<premis:eventDateTime>2006-06-06T00:00:00.001</premis:eventDateTime>
<premis:eventDetail>jhove1_1e</premis:eventDetail>
<premis:eventOutcomeInformation>
<premis:eventOutcome>successful</premis:eventOutcome>
<premis:eventOutcomeDetail>
<premis:eventOutcomeDetailNote>Well-formed and valid</premis:eventOutcomeDetailNote>
<premis:eventOutcomeDetailExtension>
<logfileInfo>
<in/>
<out/>
</logfileInfo>
</premis:eventOutcomeDetailExtension>
</premis:eventOutcomeDetail>
</premis:eventOutcomeInformation>
<premis:linkingAgentIdentifier>
<premis:linkingAgentIdentifierType>AgentID</premis:linkingAgentIdentifierType>
<premis:linkingAgentIdentifierValue>na12345</premis:linkingAgentIdentifierValue>
</premis:linkingAgentIdentifier>
<premis:linkingObjectIdentifier>
<premis:linkingObjectIdentifierType>hdl</premis:linkingObjectIdentifierType>
<premis:linkingObjectIdentifierValue>loc.music/gottlieb.09601</premis:linkingObjectIdentifierValue>
</premis:linkingObjectIdentifier>
</premis:event>

```

```

</mets:xmlData>
</mets:mdWrap>
</mets:digiprovMD>
<mets:digiprovMD ID="event2">
<mets:mdWrap MDTYPE="PREMIS:EVENT">
<mets:xmlData>
<premis:event>
<premis:eventIdentifier>
<premis:eventIdentifierType>LocalRepository</premis:eventIdentifierType>
<premis:eventIdentifierValue>E001.2</premis:eventIdentifierValue>
</premis:eventIdentifier>
<premis:eventType>ingestion</premis:eventType>
<premis:eventDateTime>2006-06-06T00:00:00.002</premis:eventDateTime>
<premis:eventDetail>ingester1_0.exe</premis:eventDetail>
<premis:eventOutcomeInformation>
<premis:eventOutcome>successful</premis:eventOutcome>
</premis:eventOutcomeInformation>
<premis:linkingAgentIdentifier>
<premis:linkingAgentIdentifierType>AgentID</premis:linkingAgentIdentifierType>
<premis:linkingAgentIdentifierValue>na12345</premis:linkingAgentIdentifierValue>
</premis:linkingAgentIdentifier>
<premis:linkingObjectIdentifier>
<premis:linkingObjectIdentifierType>hdl</premis:linkingObjectIdentifierType>
<premis:linkingObjectIdentifierValue>loc.music/gottlieb.09601</premis:linkingObjectIdentifierValue>
</premis:linkingObjectIdentifier>
</premis:event>
</mets:xmlData>
</mets:mdWrap>
</mets:digiprovMD>
<mets:digiprovMD ID="event3">
<mets:mdWrap MDTYPE="PREMIS:EVENT">
<mets:xmlData>
<premis:event>
<premis:eventIdentifier>
<premis:eventIdentifierType>LocalRepository</premis:eventIdentifierType>
<premis:eventIdentifierValue>E001.3</premis:eventIdentifierValue>
</premis:eventIdentifier>
<premis:eventType>validation</premis:eventType>
<premis:eventDateTime>2006-06-06T00:00:00.005</premis:eventDateTime>
<premis:eventDetail>jhove1_1e</premis:eventDetail>
<premis:eventOutcomeInformation>
<premis:eventOutcome>successful</premis:eventOutcome>
<premis:eventOutcomeDetail>
<premis:eventOutcomeDetailNote>Well-formed and valid</premis:eventOutcomeDetailNote>
</premis:eventOutcomeDetail>
</premis:eventOutcomeInformation>
<premis:linkingAgentIdentifier>
<premis:linkingAgentIdentifierType>AgentID</premis:linkingAgentIdentifierType>
<premis:linkingAgentIdentifierValue>na12345</premis:linkingAgentIdentifierValue>
</premis:linkingAgentIdentifier>
<premis:linkingObjectIdentifier>
<premis:linkingObjectIdentifierType>hdl</premis:linkingObjectIdentifierType>
<premis:linkingObjectIdentifierValue>loc.music/gottlieb.09602</premis:linkingObjectIdentifierValue>
</premis:linkingObjectIdentifier>
</premis:event>
</mets:xmlData>
</mets:mdWrap>
</mets:digiprovMD>
<mets:digiprovMD ID="event4">
<mets:mdWrap MDTYPE="PREMIS:EVENT">

```

```

<mets:xmlData>
<premis:event>
<premis:eventIdentifier>
<premis:eventIdentifierType>LocalRepository</premis:eventIdentifierType>
<premis:eventIdentifierValue>E001.4</premis:eventIdentifierValue>
</premis:eventIdentifier>
<premis:eventType>ingestion</premis:eventType>
<premis:eventDateTime>2006-06-06T00:00:00.006</premis:eventDateTime>
<premis:eventDetail>ingester1_0.exe</premis:eventDetail>
<premis:eventOutcomeInformation>
<premis:eventOutcome>successful</premis:eventOutcome>
</premis:eventOutcomeInformation>
<premis:linkingAgentIdentifier>
<premis:linkingAgentIdentifierType>AgentID</premis:linkingAgentIdentifierType>
<premis:linkingAgentIdentifierValue>na12345</premis:linkingAgentIdentifierValue>
</premis:linkingAgentIdentifier>
<premis:linkingObjectIdentifier>
<premis:linkingObjectIdentifierType>hdl</premis:linkingObjectIdentifierType>
<premis:linkingObjectIdentifierValue>loc.music/gottlieb.09602</premis:linkingObjectIdentifierValue>
</premis:linkingObjectIdentifier>
</premis:event>
</mets:xmlData>
</mets:mdWrap>
</mets:digiprovMD>
<mets:digiprovMD ID="event5">
<mets:mdWrap MDTYPE="PREMIS:EVENT">
<mets:xmlData>
<premis:event>
<premis:eventIdentifier>
<premis:eventIdentifierType>LocalRepository</premis:eventIdentifierType>
<premis:eventIdentifierValue>E002.1</premis:eventIdentifierValue>
</premis:eventIdentifier>
<premis:eventType>migration</premis:eventType>
<premis:eventDateTime>2006-07-06T00:00:00.006</premis:eventDateTime>
<premis:eventDetail>Adobe Photoshop</premis:eventDetail>
<premis:eventOutcomeInformation>
<premis:eventOutcome>successful</premis:eventOutcome>
</premis:eventOutcomeInformation>
<premis:linkingAgentIdentifier>
<premis:linkingAgentIdentifierType>AgentID</premis:linkingAgentIdentifierType>
<premis:linkingAgentIdentifierValue>na12345</premis:linkingAgentIdentifierValue>
</premis:linkingAgentIdentifier>
<premis:linkingObjectIdentifier>
<premis:linkingObjectIdentifierType>hdl</premis:linkingObjectIdentifierType>
<premis:linkingObjectIdentifierValue>loc.music/gottlieb.09601</premis:linkingObjectIdentifierValue>
</premis:linkingObjectIdentifier>
</premis:event>
</mets:xmlData>
</mets:mdWrap>
</mets:digiprovMD>
<mets:digiprovMD ID="event6">
<mets:mdWrap MDTYPE="PREMIS:EVENT">
<mets:xmlData>
<premis:event>
<premis:eventIdentifier>
<premis:eventIdentifierType>LocalRepository</premis:eventIdentifierType>
<premis:eventIdentifierValue>E002.2</premis:eventIdentifierValue>
</premis:eventIdentifier>
<premis:eventType>migration</premis:eventType>
<premis:eventDateTime>2007-06-06T00:00:00.006</premis:eventDateTime>

```

```

<premis:eventDetail>Adobe Photoshop</premis:eventDetail>
<premis:eventOutcomeInformation>
<premis:eventOutcome>successful</premis:eventOutcome>
</premis:eventOutcomeInformation>
<premis:linkingAgentIdentifier>
<premis:linkingAgentIdentifierType>AgentID</premis:linkingAgentIdentifierType>
<premis:linkingAgentIdentifierValue>na12345</premis:linkingAgentIdentifierValue>
</premis:linkingAgentIdentifier>
<premis:linkingObjectIdentifier>
<premis:linkingObjectIdentifierType>hdl</premis:linkingObjectIdentifierType>
<premis:linkingObjectIdentifierValue>loc.music/gottlieb.09602</premis:linkingObjectIdentifierValue>
</premis:linkingObjectIdentifier>
</premis:event>
</mets:xmlData>
</mets:mdWrap>
</mets:digiprovMD>
<mets:digiprovMD ID="agent1">
<mets:mdWrap MDTYPE="PREMIS:AGENT">
<mets:xmlData>
<premis:agent>
<premis:agentIdentifier>
<premis:agentIdentifierType>AgentID</premis:agentIdentifierType>
<premis:agentIdentifierValue>na12345</premis:agentIdentifierValue>
</premis:agentIdentifier>
<premis:agentName>LC Repository</premis:agentName>
<premis:agentType>organization</premis:agentType>
</premis:agent>
</mets:xmlData>
</mets:mdWrap>
</mets:digiprovMD>
</mets:amdSec>
<mets:fileSec>
<mets:fileGrp USE="MASTER">
<mets:file MIMETYPE="image/tiff" GROUPID="G1" ID="masterd1e102963" ADMID="object1 agent1 event1
event2 event5">
<mets:FLocat LOCTYPE="URL" xlink:href="http://lcweb2.loc.gov/natl/lib/ihas/warehouse/gottlieb/09601/ver0
1/0001.tif"/>
</mets:file>
<mets:file MIMETYPE="image/tiff" GROUPID="G1" ID="masterd1e102965" ADMID="object2 agent1 event3
event4 event6">
<mets:FLocat LOCTYPE="URL" xlink:href="http://lcweb2.loc.gov/natl/lib/ihas/warehouse/gottlieb/09601/ver0
2/0001.tif"/>
</mets:file>
</mets:fileGrp>
<mets:fileGrp USE="SERVICE">
<mets:file MIMETYPE="image/jpeg" GROUPID="G1" ID="serviced1e102963">
<mets:FLocat LOCTYPE="URL" xlink:href="http://lcweb2.loc.gov/natl/lib/ihas/service/gottlieb/09601/ver01/00
01v.jpg"/>
</mets:file>
<mets:file MIMETYPE="image/jpeg" GROUPID="G1" ID="serviced1e102965">
<mets:FLocat LOCTYPE="URL" xlink:href="http://lcweb2.loc.gov/natl/lib/ihas/service/gottlieb/09601/ver02/00
01v.jpg"/>
</mets:file>
</mets:fileGrp>
</mets:fileSec>
<mets:structMap>
<mets:div DMDID="MODS" TYPE="photo:photoObject">
<mets:div TYPE="photo:version" DMDID="ver1">
<mets:div TYPE="photo:image">
<mets:fptr FILEID="masterd1e102963"/>

```

```
<mets:fptr FILEID="serviced1e102963"/>
</mets:div>
</mets:div>
<mets:div TYPE="photo:version" DMDID="ver2">
<mets:div TYPE="photo:image">
<mets:fptr FILEID="masterd1e102965"/>
<mets:fptr FILEID="serviced1e102965"/>
</mets:div>
</mets:div>
</mets:div>
</mets:structMap>
</mets:mets>
```