

JISC



UNIVERSITY OF  
BIRMINGHAM



MIDESS Digital Preservation Requirements Specification

**Executive Summary**

The MIDESS Project is a JISC project funded under the *Digital Repositories Programme*. MIDESS explores the management of digitised content in an institutional and cross-institutional context through the development of a digital repository infrastructure. The project addresses how support can be provided for the use of digital content in a learning and research context, in an integrated manner. The partners in the project are the University of Leeds, University of Birmingham, London School of Economics (LSE) and University College London (UCL).

As part of work-package 5 of the MIDESS Project, the project manager has examined the practical considerations for undertaking long term digital preservation of the material held in the digital repository at the MIDESS institutions.

1. Introduction.....	3
2. Key Issues and Definitions .....	4
3. Repositories and Digital Preservation .....	7
3.1. File Formats .....	7
3.1.1 Raster Images .....	7
3.1.2 Raster Image formats for the MIDESS Project.....	7
3.1.3 Sound formats for the MIDESS project .....	8
3.1.4 Moving Image formats for the MIDESS project .....	8
3.2 Automatic Metadata Extraction .....	9
3.3 Digital Preservation Functionality in Repository Software Packages.....	10
3.3.1. Still Images.....	10
3.3.2 Moving Images .....	15
3.3.3 Sound.....	15
3.3.4 Metadata Schemas for Moving Images and Sound. ....	16
4. Proposals for MIDESS Partners .....	22
5. Conclusions.....	24
6. References.....	24

## **1. Introduction**

The preservation of digital material has become an important issue for organisations who wish to preserve and access their digital material long term.

Simply physically storing the digital data isn't enough for long term preservation. Details about the application used to create the digital file, rights to the digital file, ability to extract the physical attributes of the file (such as size, format, etc) and store these attributes as metadata is also important for the long term digital preservation of digital material.

This workpackage examines the key issues involved in digital preservation and how these can be practically applied to the digital repositories specialising in the storage of media such as images, sound and video.

## 2. Key Issues and Definitions

Three key types of metadata must be considered in the preservation of digital objects. These are **Technical Metadata**, **Administrative Metadata** and **Discovery Metadata**.

- **Technical Metadata** describes the physical attributes of digital objects.
- **Administrative Metadata** describes the rights ownership and provenance.
- **Discovery Metadata** describes how to locate access and use digital content in the long term.

According to the LIFE project<sup>1</sup> key factors in the preservation of digital objects include:

1. Frequency of action – how often preservation action needs to be taken to preserve the digital material long term.  
  
If preservation action is not taken, then files in specific formats may be unreadable. For example files originally written in Ventura Publisher are now extremely difficult to read since the software has been discontinued.
2. Technology watch – identifying the points at which preservation actions need to occur.  
  
This is the time spent evaluating the current technology and software to determine whether files need to be migrated to a more suitable format.
3. Availability of digital preservation software tools – when the tools are available and the time for these new digital preservation software tools to be developed.  
  
The LIFE project predicts that the number of digital preservation tools will increase significantly and that about 90% of digital preservation needs will be met in 20 years time.
4. Cost of tools – cost of developing a solution.  
  
This is usually the cost of hiring a programmer and developing a practical software tool for digital preservation.
5. Complexity of file formats – how the file format itself affects the cost of a preservation action.  
  
Aspects of a file format that impact on the cost of preserving objects include size, complexity and whether it is open or proprietary. For example on a scale from 0 to 1 for complexity, formats such as ASCII rank as 0 and large Oracle database files rank as 1.
6. Preservation strategies – how the model addresses the use of different approaches to preserving digital objects.  
  
For example should migration or emulation be used to access the preserved files in the long term?
7. Quality assurance – checking the accuracy and effectiveness of a preservation action.

The simplest method of digital preservation is bitstream preservation<sup>2</sup> where the original binary digits are preserved in an uninterrupted state. Commonly called “backing up your data”

## MIDESS: WP 5 – Digital preservation requirements specification

this is the minimum strategy that should be employed for all digital preservation. There are a number of other approaches to the digital preservation issue which can be broadly subdivided into two methods:

- Either at some stage either old data must be altered in order to operate in a new technical environment (using such techniques as migration or format standardisation of the original material etc).

or alternatively:

- The new environment must be modified so that it can handle the old data (emulation, virtual computers etc).

### Migration

Migration of digital material is the potential ability to transfer the material into a new digital format without losing any of the digital content within the original migrated file. Thus there is a transfer of the file from one hardware/software configuration to another (updated) hardware/software configuration. A major issue with migration is that there is the potential for information to be lost in the migration. Also since the resultant migrated file is not an exact copy of the original file, then to what extent this technique of migration can be said to meet the requirements of digital preservation is open to debate. Systems are being developed which periodically migrate files from their original formats to more readable formats automatically. These systems store the new migrated file and either retain the original file or discard it.

### Format Standardisation

Format standardisation is the acceptance of a standard format which is supported (and which will continue to be supported) for the foreseeable future. Supporting a large variety of formats in a digital repository may require duplication of digital files into multiple formats or users will not be able to view the files if they are in uncommon formats without access to the associated applications. Many repositories now standardise on specific supported formats when ingested into a repository. Increasingly there are systems which on ingest automatically convert files to the appropriate standard formats for that repository. Proprietary formats (such as Microsoft Word) are usually avoided in favour of more open formats (such as PDF). Avoiding proprietary formats helps to ensure that should the format be discontinued then continued access will still be available provided that the file is contained in a non-proprietary format.

### Emulation

Emulation can be considered to be the development of software/hardware combinations to replicate the behaviour of obsolete processes. These hardware processes may include systems such as interfaces, operating systems or hardware configurations. By replicating the behaviour of obsolete processes this enables digital material stored using these obsolete processes to be accessed by modern systems.

### Virtual Computers

The Universal Computer is a form of emulation and is a virtual representation of a simplified computer that will run on any existing hardware platform. Software emulates the original obsolete software/hardware combination. The major problem here is that a Universal Virtual Computer emulator will need to be developed for every hardware/software configuration on which the file will be accessed throughout its useable life, therefore the problem moves from that of digital preservation to the continued development of developing suitable emulators for hardware/software emulations.

**The Open Archival Information System reference model. (OAIS)** is a best practice reference model for long term preservation and is an ISO standard.

Extensive details regarding the OAIS standard have been detailed elsewhere <sup>3</sup>

## MIDESS: WP 5 – Digital preservation requirements specification

However the standard includes the following mandatory responsibilities:

1. Negotiate for and accept appropriate information from information producers
2. Obtain sufficient control of the information in order to meet long-term preservation objectives.
3. Determine the scope of the archive's user community
4. Ensure that the preserved information is independently understandable to the user community, in the sense that the information can be understood by users without the assistance of the information producer.
5. Follow documented policies and procedures to ensure that information is preserved against all reasonable contingencies and to enable dissemination of authenticated copies of the preserved information in its original form, or in a form traceable to the original.
6. Make the preserved information available to the user community.

OAIS provides a framework for capturing information packages in a way that they may be stored (or migrated, or emulated) according to two layers of metadata. In order to keep metadata complexity low, any archive following OAIS is asked to implement a dependency classification system for structural, semantic and representational metadata.

The recent JISC-funded Feasibility and Requirements Study for Preservation of E-Prints<sup>4</sup> argued that there is a unique window of opportunity to address the preservation requirements of repositories at the beginning of their adoption rather than leaving it until the lack of preservation management becomes an issue and content is no longer accessible. A key recommendation of this report was the establishment of a repository infrastructure based upon the OAIS reference model.

## **3. Repositories and Digital Preservation**

For the purposes of the MIDESS project we need to specifically examine the preservation issues for multimedia material such as images, moving images and sound.

Specifically we need to address what metadata elements should be included for the practical preservation of metadata for digitised multimedia material. In practise these metadata elements equate to database fields storing textual descriptions (metadata) in a relational database stored alongside the actual digital objects.

### **3.1. File Formats**

#### **3.1.1 Raster Images**

At present the most reliable form of preservation for raster images is considered to be the format migration approach. The AHDS report<sup>5</sup> strongly recommends that different versions of the images are kept, those in their original format and those migrated to a new format. This ensures that when other preservation methods (such as emulation or the universal virtual computing methods) become more suitable as potential preservation techniques, then these techniques can be applied to the original source material in combination with, or as an alternative to the format migration approach.

The AHDS report recommends that uncompressed TIFF version 6 is considered the most appropriate format to use for long term preservation of images. If the preservation metadata is in the form of TIFF formatted files and these files have been saved in a compressed TIFF format, then these compressed TIFF files should be opened with an application such as Adobe Photoshop and then saved as uncompressed TIFF files version 6 for digital preservation.

In addition to Uncompressed TIFF, there is however other emerging formats such as JPEG2000, PNG and DNG which should be watched to see if these become suitable candidates for long term digital preservation. JPEG2000 has a mechanism that allows metadata to be embedded as XML within its actual image file. Thus copying or moving the JPEG2000 files automatically takes with it the accompanying metadata attached to the file. This metadata contained with the file can be wide-ranging and can include content description metadata such as keywords (i.e. the "who", "what", "when" and "where" aspects of the image).

#### **3.1.2 Raster Image formats for the MIDESS Project**

It has been decided to adopt a practical approach to the digital format issue for files held with the MIDESS digital repository's and we intent to adopt the following solutions:-

For images, it is proposed that uncompressed TIFF version 6 format in addition to a web enabled format such as JPG will be the formats of choice for the digital repository. Here the uncompressed TIFF version 6 TIFF file is used for the long term preservation of the digital object and the jpg file is used for display within a web browser.

Unfortunately as part of the MIDESS project we have found that there are a large number of image files which have already been saved in a variety of image formats (such as BMP, PICT etc). In these situations it has been decided for the MIDESS project that these files will be

## **MIDESS: WP 5 – Digital preservation requirements specification**

preserved and the files will also be converted into web enabled formats such as JPG where possible, for easy display within the digital repository.

Under these circumstances while it has been felt important to convert original digital images into a format which can be easily viewed via the web/digital repository, it has also been decided to preserve the original files in their original formats supplied by the depositors. It was felt that while converting the file into a web enabled format would probably be the most benefit to the majority of people viewing the file via the digital repository, this may not be the case in every instance. Throwing away the original source file simply because it wasn't in a long term preservation format (such as the widely recognised long term image preservation format uncompressed TIFF) was felt to be inappropriate. MIDESS feels it is important however, to ensure that the depositors are aware that long term preservation may have been compromised because the original file owner had not saved the digital material in an appropriate digital preservation format in the first place.

### **3.1.3 Sound formats for the MIDESS project**

MIDESS workpackage 3 showed that both the MP3 and WAV format were by far the most popular formats.

The AHDS Digital Imaging Archive Study conducted by the AHDS<sup>6</sup> suggest that either uncompressed broadcast wave format (BWF) or AIFF, using Linear PCM as the encoding method are used for the preservation for audio files.

The MIDESS project has decided to recommend the digitisation of sound files in MP3 and AIFF formats where possible for storage in the digital repository. MP3 format as the primary playback format and AIFF as the most appropriate preservation format. This assumes that the digitised sound files do not already exist and that the sound files are being created. Where the sound files already exist then the sound file (in the format provided to the repository) will be saved in the digital repository irrespective of its format. This file will also be converted to a streamed MP3 format for playback directly from our University of Leeds streamed server wherever possible.

### **3.1.4 Moving Image formats for the MIDESS project**

Preservation of digital moving images for the foreseeable future will require the adoption of a migration approach<sup>8</sup>. If lossy compression is involved signal quality will decrease with each migration. In general therefore moving images should be saved in an uncompressed format. Dissemination/distribution versions can be lossy, compression is preferred because of the benefit of the reduction in bandwidth achieved.

MIDESS workpackage 3 demonstrated that no one moving image format stood out at the most popular format – WMV, AVI, MPEG, QuickTime and Real Player all were equal in popularity with people creating moving images (video). It was felt that QuickTime was probably the most compatible across platforms, however in special circumstances such as at the University of Leeds other factors have a strong contributing factor.

Currently the University of Leeds has a streaming media server which can natively support Microsoft's WMV files. This streaming server<sup>7</sup> at the University of Leeds can run WMV files streamed on the server at different Image size and bit rate settings. This streaming server service will shortly be adapted to stream both QuickTime and Real Media files as well as Microsoft Windows WMV files. Digital files on this streaming server can be accessed directly from within the University of Leeds digital repository by anybody who has the necessary rights to access the files. For this reason the University of Leeds will be creating files in WMV format to run on the streaming server.

## **MIDESS: WP 5 – Digital preservation requirements specification**

Where the moving image files do not already exist (not already digitised), then for long term preservation of moving images, Leeds expects to use the uncompressed or lossless compression (motion) jpeg 2000 format inside a JPEG2000 wrapper as a long term preservation format<sup>8</sup>. Other preferred wrapper formats are AVI, QuickTime and WMV as long as the video and audio bitstreams within the wrappers are uncompressed.

However where the moving image file already exists and there is no possibility of converting the original video/film into a suitable format, then MIDESS will preserve the file provided irrespective of its format. It will additionally convert the file into a streamed file for storage and streamed playback from within the digital repository where possible.

### **3.2 Automatic Metadata Extraction**

Currently metadata is primarily entered manually in many repositories, either directly, or via a batch process where the metadata is entered into a spreadsheet or database and then the contents of the resulting file imported directly into a repository. Metadata extraction tools are being developed which can automatically extract metadata (such as file formats) from digital objects.

Examples of these metadata extraction tools include:

Droid<sup>9</sup>

National Library of New Zealand Metadata Extraction Tool<sup>10</sup>

Many digital repositories display metadata information such as file name, file size, file extension etc, however none of the digital repositories that the author is aware of have the ability to extract and seamlessly store this extracted metadata as elements of a preservation metadata schema directly within the respective repository.

### 3.3 Digital Preservation Functionality in Repository Software Packages.

#### 3.3.1. Still Images

It has been suggested by AHDS<sup>11</sup> that a minimum preservation metadata element set based upon the PREMIS data dictionary, and thus a mandatory number of elements (database fields) can be used from the PREMIS metadata schema to adequately describe digital objects. Thus rather than including all the metadata elements from the PREMIS metadata schema in a repository to describe a digital object, including only most important (mandatory) elements from the PREMIS metadata schema is a much more practical solution.

It has also further been suggested by AHDS<sup>12</sup> that the metadata that can be used for the practical application of digital preservation should consist of a combination of descriptive metadata schemas such as **Simple Dublin Core** for resource discovery elements along with elements from PREMIS for management and technical metadata, and some specific elements from NISO Z39.87 metadata schema.

**PREMIS**<sup>13</sup> provides a set of core metadata elements that are needed to support the preservation of all kinds of digital resources, regardless of their data type (i.e. not just digital images). It covers the fundamental entities Objects, Events, Agents, and Rights that are common to all resources. It is therefore very relevant although it doesn't cover all of the specific technical aspects relating to a particular data type (such as digital images).

**NISO Z39.87**<sup>14</sup> adopts many of the basic metadata elements from PREMIS but then adds a large number of technical metadata elements (relating to image capture, image assessment and change history) that are specific for the management and preservation of raster images. It has an XML schema (MIX<sup>15</sup>) which is an extension of the METS schema<sup>16</sup> and stands every chance of widespread uptake and long-term support because it has the support of OCLC, RLG and the Library of Congress.

Practically the AHDS has suggested that suitable fields for inclusion in a digital repository which would enable the repository to act as a suitable vessel for long term digital preservation would be the inclusion of the following metadata elements (database fields).

Element	Metadata Standard	Definition	Comment
Title	Dublin Core	The name given to a resource	Usually the Title will be a name by which the resource is formally known
Creator	Dublin Core	An Entity primarily responsible for making the content of the resource	Examples include A person, organisation or service.
Subject	Dublin Core	The topic of the content of the resource	Often expressed as keywords that describe the entity – best from a controlled vocabulary
Date	Dublin Core	A date associated with an event in the life cycle of the resource	Ideally should follow the ISO 8601 <sup>1</sup> and should be in the format YYYY-MM-DD
Identifier	Dublin Core	An unambiguous reference to the	Recommended to identify the resource

## MIDESS: WP 5 – Digital preservation requirements specification

		resource within a given context	by a string or number confirming to a formal identification system.
Provenance	Dublin Core	A statement of changes of ownership and custody of the resource since its creation that is significant for its authenticity and interpretation	Descriptions of changes successive custodians have made to the resource.
Rights	Dublin Core	Information about rights held in and over the resource	Typically a rights management statement for the resource IPR, Copyright, and various property rights.

**The Object Identifier Definition** – The Object Identifier is used to uniquely identify the object within the preservation repository system in which it is stored. Each data object held in the preservation repository must have a unique identifier to relate it to descriptive, technical, and other metadata. An identifier may be created by the repository system at the time of ingest or it may be created or assigned outside of the repository and submitted with an object as metadata. Identifiers can be automatically or manually generated. Recommended practice is for repositories to use identifiers automatically created by the repository as the primary identifier in order to ensure that identifiers are unique and usable by the repository.

The **ObjectIdentifierType** and **ObjectIdentifierValue** elements which together make up the **Object Identifier**.

Element	Metadata Standard	Definition	Comment
<b>ObjectIdentifierType</b>	PREMIS	A designation of the domain within which the object identifier is unique	Identifier values cannot be assumed to be unique across domains; the combination of objectIdentifierType and objectIdentifierValue should ensure uniqueness. Value should be taken from a controlled vocabulary
<b>ObjectIdentifierValue</b>	PREMIS	A designation used to uniquely identify the object within the preservation system in which it is stored	The value of the objectIdentifier

## MIDESS: WP 5 – Digital preservation requirements specification

### The PreservationLevel Element

**PreservationLevel Definition.** A value indicating the set of preservation functions expected to be applied to the object. Examples include STORE, MIXED, FULL etc.

Element	Metadata Standard	Definition	Comment
PreservationLevel	PREMIS	A value indicating the set of preservation functions expected to be applied to the object	Some preservation repositories will offer multiple preservation options depending on factors such the value or uniqueness of the material, the “preservability” of the format, the amount the customer is willing to pay, etc. Its value should where possible be taken from a controlled vocabulary.

### The ObjectCategory Element

Examples include representation, file, bitstream. A filestream should be considered a file.

Element	Metadata Standard	Definition	Comment
ObjectCategory	PREMIS	The category of object to which the metadata applies	Preservation repositories are likely to treat different categories of objects (representations, files, and bitstreams) differently in terms of metadata and data management functions. Value should be taken from a controlled vocabulary.

### The MessageDigestAlgorithm and MessageDigest Elements.

Both **MessageDigestAlgorithm** and **MessageDigest** are fields which are part of the checks that are used to ensure that the particular content Information object has not been altered in an undocumented or unauthorized way (commonly called ‘fixity’) The **MessageDigestAlgorithm** is the specific algorithm used to construct the message digest for the digital object. Examples include cryptographic hash functions such as HAVAL and SHA-1 where a variable length string is converted into a string fixed in length.

Element	Metadata Standard	Definition	Comment
MessageDigestAlgorithm	PREMIS	The specific algorithm used to construct the message digest from the digital object	Value should be taken from a controlled vocabulary
MessageDigest	PREMIS	The output of the message digest algorithm	This must be stored so that it can be compared in future fixity checks.
FormatName	PREMIS	TIFF, JPEG etc.	Value should be taken from

**MIDESS: WP 5 – Digital preservation requirements specification**

			a controlled vocabulary.
<b>FormatVersion</b>	PREMIS	The version of the format (2.0, 3.1 etc)	Many authority lists of format names are not granular enough to indicate version of example, MIME Media types.
<b>StorageMedium</b>	PREMIS	The physical medium on which the object is stored (e.g. magnetic tape, hard disk, CD-ROM, DVD etc)	The repository needs to know the medium on which an object is stored in order to know how and when to do media refreshment and media migration.
<b>EventIdentifierType</b>	PREMIS	A designation of the domain within which the event identifier is unique.	For most preservation repositories, the eventIdentifierType will be their own internal numbering system. It can be implicit within the system and provided explicitly only if the data is exported.
<b>EventIdentifierValue</b>	PREMIS	The value of the eventIdentifier	
<b>EventType</b>	PREMIS	A categorisation of the nature of the event.	Categorizing events will aid the preservation repository in machine processing of event information, particularly in reporting. Value should be taken from a controlled vocabulary.
<b>EventDateTime</b>	PREMIS	The single date and time or date and time range at or during which the event occurred.	Any date/time convention may be used, as long as it is consistent and can be translated into ISO 8601 for export if necessary.
<b>AgentIdentifierType</b>	PREMIS	A designation of the domain in which the agent identifier is unique.	Value should be taken from a controlled vocabulary
<b>AgentIdentifierValue</b>	PREMIS	The value of the agentIdentifier.	May be a unique key or a controlled textual form of name.
<b>ColorSpace</b>	NISO Z39.87	A designation of the colour model of the decompressed image data.	Commonly called colour spaces, these colour models (eg RGB, CMYK) are drawn from common file formats used to render digital still images. Some colour models may be pertinent to certain file types (e.g. TIFF) while others are more device dependent or independent (calibrated) colour models. ColorSpace should be a text description.
<b>ImageWidth</b>	NISO z39.87	A specification of the width of the	The image width may be the shorter or longer

**MIDESS: WP 5 – Digital preservation requirements specification**

		digital image, i.e. horizontal or X dimension pixels	dimension of the image, depending on the orientation of the camera or scanner during image capture. For multiple resolution image file formats, value shall specify the highest resolution.
<b>ImageHeight</b>	NISO z39.87	A specification of the height of the digital image, i.e. vertical or Y dimension in pixels	The image height may be the shorter or longer dimension of the image, depending upon the orientation of the camera or scanner during image capture. For multiple-resolution image file formats, value shall specify the highest resolution.
<b>BitsPerSample</b>	NISO Z39.87	The number of bits per component for each pixel	This field is used to describe the number of bits for each sample (or channel) expressed in the same order given in colorSpace. BitsPerSample is equivalent to bit depth. It gives the sample rate per colour channel – so, for instance, 8,8,8 is 24bit.

PREMIS and NISO Z39.87 preservation metadata schemas complement each other well because they focus on different parts of digital preservation. NISO Z29.87 provides more comprehensive technical metadata than PREMIS while for rights, authenticity and provenance, elements within PREMIS are more comprehensive than those in NISO Z39.87.

### 3.3.2 Moving Images

The following requirements are recommended for moving images<sup>18</sup>:

- Larger picture size preferred over smaller picture size.
- Content from high definition sources preferred over content from standard definition, assuming picture size is equal or greater.
- Encodings that maintain frame integrity preferred over formats that use temporal compression.
- Uncompressed or lossless compressed preferred over lossy compressed.
- Higher bit rate (often expressed as *megabits per second*) preferred over lower for same lossy compression scheme.
- Extended dynamic range (scene brightness) preferred over "normal" dynamic range for such items as Digital Cinema or scanned motion picture film.
- Surround sound encoding only necessary if essential to creator's intent. In other cases, stereo or monaural sound is preferred.

Practically this means the use of uncompressed or lossless compression (motion) jpeg2000 format inside a JPEG2000 wrapper for moving image files. For commercial movies the use of the DCDM format is best.

### 3.3.3 Sound

For sound the general requirements are:

- Higher sampling rate (usually expressed as kHz) preferred over lower sampling rate.
- 24-bit sample word-length preferred over shorter
- Linear PCM (uncompressed)<sup>19</sup> preferred over compressed (lossy or lossless)
- Higher data rate (e.g. 128 kilobits per second) preferred over lower data rate for same compression scheme and sampling rate.
- Advanced Audio Coding (AAC)<sup>20</sup> compression preferred over MPEG-layer 3 (MP3)<sup>21</sup>
- Surround sound (5.1 or 7.1) encoding only necessary if essential to creator's intent. In other cases, uncompressed encoding in stereo is preferred

In practice this means the use of either uncompressed broadcast wave format (BWF<sup>22</sup>) or AIFF<sup>23</sup>, using Linear PCM as the encoding method, for audio files.

**3.3.4 Metadata Schemas for Moving Images and Sound.**

Simple Dublin Core is probably the best basic standard to develop a minimum applicable element set for both moving images and sound resources<sup>18</sup>. Title, Creator, Subject, Date, Identifier and Rights are likely to be essential for all moving image and audio types of resource. However it should be noted that Dublin Core has no built-in mechanisms for making fine distinctions between different kinds of entry in each of these elements. Generally speaking is suitable or metadata exchange and harvesting of the resource at the item or digital object level. The AHDS<sup>18</sup> have suggested a metadata schema which also includes additional elements from other schemas such as PREMIS, VideoMD, SoundMD and MPEG-7.

In Addition to those elements from PREMIS contained in the image metadata schema the suggested metadata Schemas for moving images and sound include elements from the AudioMD (Audio Technical Metadata Schema) and VideoMD (Video Technical Metadata Schema) and MPEG-7

The Library of Congress Audio-Visual Prototyping Project has produced a draft METS extension metadata schema and data dictionary for AudioMD audio resources:

Draft data dictionary: [http://www.loc.gov/rr/mopic/avprot/DD\\_AMD.html](http://www.loc.gov/rr/mopic/avprot/DD_AMD.html)

Draft schema: [http://www.loc.gov/rr/mopic/avprot/AMD\\_020409.xsd](http://www.loc.gov/rr/mopic/avprot/AMD_020409.xsd)

AudioMD contains 37 technical metadata elements for describing an audio object.

The Library of Congress Audio-Visual Prototyping Project also produced VideoMD, a draft METS extension metadata schema and data dictionary for video objects that contains 16 technical metadata elements.

Draft data dictionary: [http://www.loc.gov/rr/mopic/avprot/DD\\_VMD.html](http://www.loc.gov/rr/mopic/avprot/DD_VMD.html)

Draft schema: <http://lcweb-2.loc.gov/mets/Schemas/VMD.xsd>

MPEG-7 provides many descriptors for the searching and filtering of specific types of content. For example for the audio content there is waveform, fundamental frequency, spoken content, timbre, etc. For visual content there is colour, texture, shape, motion, etc.

**Elements used in both Moving Image and Audio Schemas.**

<b>Element</b>	<b>Metadata Standard</b>	<b>Definition</b>	<b>Comment</b>
<b>Title</b>	Dublin Core	The name given to a resource	Usually the Title will be a name by which the resource is formally known
<b>Creator</b>	Dublin Core	An Entity primarily responsible for making the content of the resource	Examples include A person, organisation or service.
<b>Subject</b>	Dublin Core	The topic of the content of the resource	Often expressed as keywords that describe the entity – best from a controlled vocabulary
<b>Date</b>	Dublin Core	A date associated with an event in the	Ideally should follow the ISO 8601 <sup>1</sup>

**MIDESS: WP 5 – Digital preservation requirements specification**

		life cycle of the resource	and should be in the format YYYY-MM-DD
<b>Identifier</b>	Dublin Core	An unambiguous reference to the resource within a given context	Recommended to identify the resource by a string or number conforming to a formal identification system.
<b>Provenance</b>	Dublin Core	A statement of changes of ownership and custody of the resource since its creation that is significant for its authenticity and interpretation	Descriptions of changes successive custodians have made to the resource.
<b>Rights</b>	Dublin Core	Information about rights held in and over the resource	Typically a rights management statement for the resource IPR, Copyright, and various property rights.

**The Object Identifier Definition** – The Object Identifier is used to uniquely identify the object within the preservation repository system in which it is stored. Each data object held in the preservation repository must have a unique identifier to relate it to descriptive, technical, and other metadata. An identifier may be created by the repository system at the time of ingest or it may be created or assigned outside of the repository and submitted with an object as metadata. Identifiers can be automatically or manually generated. Recommended practice is for repositories to use identifiers automatically created by the repository as the primary identifier in order to ensure that identifiers are unique and usable by the repository.

The **ObjectIdentifierType** and **ObjectIdentifierValue** elements which together make up the **Object Identifier**.

Element	Metadata Standard	Definition	Comment
<b>ObjectIdentifierType</b>	PREMIS	A designation of the domain within which the object identifier is unique	Identifier values cannot be assumed to be unique across domains; the combination of objectIdentifierType and objectIdentifierValue should ensure uniqueness. Value should be taken from a controlled vocabulary
<b>ObjectIdentifierValue</b>	PREMIS	A designation used to uniquely identify the object within the preservation system in which it is stored	The value of the objectIdentifier

## MIDESS: WP 5 – Digital preservation requirements specification

### The PreservationLevel Element

**PreservationLevel Definition.** A value indicating the set of preservation functions expected to be applied to the object. Examples include STORE, MIXED, FULL etc.

Element	Metadata Standard	Definition	Comment
PreservationLevel	PREMIS	A value indicating the set of preservation functions expected to be applied to the object	Some preservation repositories will offer multiple preservation options depending on factors such the value or uniqueness of the material, the “preservability” of the format, the amount the customer is willing to pay, etc. Its value should where possible be taken from a controlled vocabulary.

### The ObjectCategory Element

Examples include representation, file, bitstream. A filestream should be considered a file.

Element	Metadata Standard	Definition	Comment
ObjectCategory	PREMIS	The category of object to which the metadata applies	Preservation repositories are likely to treat different categories of objects (representations, files, and bitstreams) differently in terms of metadata and data management functions. Value should be taken from a controlled vocabulary.

### The MessageDigestAlgorithm and MessageDigest Elements.

Both **MessageDigestAlgorithm** and **MessageDigest** are fields which are part of the checks that are used to ensure that the particular content Information object has not been altered in an undocumented or unauthorized way (commonly called ‘fixity’) The **MessageDigestAlgorithm** is the specific algorithm used to construct the message digest for the digital object. Examples include cryptographic hash functions such as HAVAL and SHA-1 where a variable length string is converted into a string fixed in length.

Element	Metadata Standard	Definition	Comment
MessageDigestAlgorithm	PREMIS	The specific algorithm used to construct the message digest fro the digital object	Value should be taken from a controlled vocabulary
MessageDigest	PREMIS	The output of the message digest algorithm	This must be stored so that it can be compared in future fixity checks.

**MIDESS: WP 5 – Digital preservation requirements specification**

<b>FormatName</b>	PREMIS	TIFF, JPEG etc.	Value should be taken from a controlled vocabulary.
<b>FormatVersion</b>	PREMIS	The version of the format (2.0, 3.1 etc)	Many authority lists of format names are not granular enough to indicate version of example, MIME Media types.
<b>StorageMedium</b>	PREMIS	The physical medium on which the object is stored (e.g. magnetic tape, hard disk, CD-ROM, DVD etc)	The repository needs to know the medium on which an object is stored in order to know how and when to do media refreshment and media migration.
<b>EventIdentifierType</b>	PREMIS	A designation of the domain within which the event identifier is unique.	For most preservation repositories, the eventIdentifierType will be their own internal numbering system. It can be implicit within the system and provided explicitly only if the data is exported.
<b>EventIdentifierValue</b>	PREMIS	The value of the eventIdentifier	
<b>EventType</b>	PREMIS	A categorisation of the nature of the event.	Categorizing events will aid the preservation repository in machine processing of event information, particularly in reporting. Value should be taken from a controlled vocabulary.
<b>EventDateTime</b>	PREMIS	The single date and time or date and time range at or during which the event occurred.	Any date/time convention may be used, as long as it is consistent and can be translated into ISO 8601 for export if necessary.
<b>AgentIdentifierType</b>	PREMIS	A designation of the domain in which the agent identifier is unique.	Value should be taken from a controlled vocabulary
<b>AgentIdentifierValue</b>	PREMIS	The value of the agentIdentifier.	May be a unique key or a controlled textual form of name.
<b>MediaFormat/BitRate</b>	MPEG-7	Indicates the nominal bit rate in bits per second or kbps of the audio or video instance.	Eg 64, 128, 256, etc. Maximum and Minimum should be used to record the minimum and maximum numerical value for the BitRate in cases of variable bitrate.
<b>MediaInformation/MediaProfile/</b>	MPEG-7	Duration the	The ISO 8601 syntax

**MIDESS: WP 5 – Digital preservation requirements specification**

<b>MediaFormat/Duration</b>		Audio or Video content. The elapsed time of the entire file.	should be used. <a href="http://www.w3.org/TR/NOTE-datetime">www.w3.org/TR/NOTE-datetime</a>
<b>Size</b>	PREMIS	Indicates the size, in bytes, of the file where the video instance is stored.	Optional

Essential additional elements for moving images

<b>Essential additional elements for moving images</b>			
<b>VisualCoding/Frame/@aspectRatio</b>	MPEG-7	The desired aspect ratio of the image on Screen	Eg 4:3
<b>VisualCoding/Pixel/@bitsPer</b>	MPEG-7	The number of bits of sample depth	Eg 8,24 etc
<b>VisualCoding/Format/Frame @width</b>	MPEG-7	The number of frames per second at which the video source was digitised	Eg Frame rate = 25
<b>VisualCoding/Format/Frame/@width</b>	MPEG-7	The horizontal size of the video frame measured by number of pixels	Frame Width = 360
<b>VisualCoding/Format/Frame/@height</b>	MPEG-7	The vertical size of the video frame measured by number of pixels	Frame Height = 240
<b>VisualCoding/Pixel@resolution</b>	MPEG-7	Resolution of digital video source item expressed as horizontal lines	
<b>Sampling</b>	VideoMD	The video sampling format (in terms of luminance and chrominance)	Eg 4:2:0, 4:2:2 2:4:4
<b>Scan</b>	TV-Anytime	An indication whether the digital video item is scanned in an interlaced or progressive	Interlaced or Progressive

**MIDESS: WP 5 – Digital preservation requirements specification**

		mode.	
<b>VisualCoding/Format/Name</b>	MPEG-7	The encoding method of the visual component of a resource. (the codec)	MPEG-1 Video etc
<b>Sound</b>	VideoMD	Indication of the presence of sound in the video file	Yes or No

Essential additional elements for sound

<b>Essential additional elements for sound</b>			
<b>AudioCoding/Format/Name</b>	MPEG-7	The encoding method of the sound component of the sound (codec)	
<b>AudioCoding/AudioChannels</b>	MPEG-7	The number of channels of audio	Eg 1,2,3,4,5
<b>AudioCoding/bitPer</b>	MPEG-7	Number of bits per audio	Eg 16,20,24
<b>Sampling_frequency</b>	audioMD	Rate at which the audio was sampled	Expressed in kHz 22,44.1, 48,96
<b>Audio_block_size</b>	audioMD	Size of an audio block in bytes	
<b>First_sample_offset</b>	audioMD	Location of the first valid sound byte in the file	
<b>First_valid_byte_block</b>	audioMD	Location of the first valid sound byte in the block.	
<b>Last_valid_byte_block</b>	audioMD	Location of the last valid sound byte in the block	

## 4. Proposals for MIDESS Partners

Currently the digital repository at the University of Leeds (which uses Endeavor's Curator digital repository software) does not specifically support digital preservation features such as automatic conversion of formats, emulation, workflow etc.

Version 4.0 of the Endeavor Curator software has however introduced the ability to add technical metadata although the elements available are somewhat limited and are only available to users who have access to either the staff client or system admin client.

This is shown below:-

**Add Digital Resource**

Object Source

Web Link    Local File    Repository File    Windows Clipboard

URL:  ...

Object Type:

Type Category:

Technical Metadata:

Property Name	Property Value
bit_depth	
color_space	
compression_codec	

    

This figure shows that bit\_depth, color\_space and compression\_codec have been included in the technical metadata however these are required to be entered by the user rather than automatically extracted by the system.

A new solution by Endeavor called Kronos has been proposed for the collection, management and long-term preservation and access of digital content for the Curator digital repository software. This was announced via the National Library of New Zealand's website<sup>25</sup> at on 7<sup>th</sup> August 2006.

It is proposed that Kronos will support features that include submission tracking, easy metadata management and access/dissemination workflows and detailed digital rights management. Kronos will also offer the ability to migrate the files from outdated formats to more appropriate formats. One of the principals behind Kronos is that preservation and access belong together –rather than being mutually exclusive concepts. Thus preservation support will be built directly into the Kronos system rather than as an add-on feature. Kronos will apply the Open Archival Information System (OAIS) model<sup>26</sup>.

The Kronos project is intended to fully support the OAIS model and provide software which can be used with all Curator digital repository software no matter how they are currently

## MIDESS: WP 5 – Digital preservation requirements specification

configured. Endeavor is working with the National Library of New Zealand to build and implement Kronos. It is expected that Kronos will be beta tested in early 2007 with a currently proposed release date of mid-late 2007.

The University of Birmingham as part of the MIDESS project uses Dspace. Dspace was one of the earliest repository systems to consider some of the basic issues of preservation, DSpace captures details of the specific file formats users submit and maintains a bitstream format for each bitstream in the system. System administrators can maintain a registry of known bitstream formats and the preservation service level available for each format type; however, if the format of the bitstream is unknown, the system will not be able to reliably support preservation and future access or re-use of the file contents. Most Dspace Repositories maintain lists of 'supported' and 'unsupported' file formats<sup>26</sup>. The University of Birmingham does not intend to specifically implement digital preservation features as part of the MIDESS project.

LSE's uses Fedora digital repository open source software and this software has a strong component of digital preservation built in to its data model. Exactly how digital preservation is built into Fedora is covered in detail in the article "Digital Preservation"<sup>27</sup>. Fedora uses concepts such as "encapsulation" which is simply a way to group together all the relevant material for the digital object and to manage the resulting digital object as one. Other features within Fedora that support digital preservation include persistent identifiers and detailed audit trails. Within audit trails for example each change to an object results, upon saving in a new date/time stamp being created to record the modification. LSE intends to make full use of the features provided within Fedora once they are in a position to do so.

## 5. Conclusions.

The author considers that digital preservation is an important issue for digital repositories that continues to be addressed; however practical techniques which implement these standards lags behind the theoretical standards currently underlining digital preservation requirements.

While models such as OAIS have been discussed at some length, practical techniques for implementing these standards continue to remain at the relatively early stages.

The University of Leeds has decided to implement a subset of PREMIS and NISO Z29.87 as an addition to the descriptive metadata schemas (Qualified Dublin Core, EAD, METS) supported at the University of Leeds digital repository. Once the Kronos system becomes commercially available we will evaluate this with a view to its adoption. Further information will be provided about this in MIDESS Workpackage 4 Metadata.

Given the variety of media formats for images, sound and video, the University of Leeds has tried to remain flexible in the formats of the material it supports.

## 6. References.

1. LIFE: Life Cycle Information for E-Literature. <http://www.ucl.ac.uk/lslifeproject>
2. Digital Images Archiving Study. Arts and Humanities Data Service (AHDS). [http://www.jisc.ac.uk/uploaded\\_documents/FinaldraftImagesArchivingStudy.pdf#search=%22FinaldraftImagesArchivingStudy.pdf%22](http://www.jisc.ac.uk/uploaded_documents/FinaldraftImagesArchivingStudy.pdf#search=%22FinaldraftImagesArchivingStudy.pdf%22)
3. SHERPA DP: Creating A Persistent Preservation Environment For Institutional Repositories. Arts and Humanities Data Service (AHDS). <http://ahds.ac.uk/about/projects/sherpa-dp/>
4. Focus on access to Institutional Repositories (FAIR) [http://www.jisc.ac.uk/whatwedo/programmes/programme\\_fair.aspx](http://www.jisc.ac.uk/whatwedo/programmes/programme_fair.aspx)
5. Digital Images Archiving Study. Arts and Humanities Data Service (AHDS). [http://www.jisc.ac.uk/uploaded\\_documents/FinaldraftImagesArchivingStudy.pdf#search=%22FinaldraftImagesArchivingStudy.pdf%22](http://www.jisc.ac.uk/uploaded_documents/FinaldraftImagesArchivingStudy.pdf#search=%22FinaldraftImagesArchivingStudy.pdf%22)
6. Digital Images Archiving Study. Arts and Humanities Data Service (AHDS). (<http://ahds.ac.uk/about/projects/archiving-studies/index.htm>)
7. Digital Delivery of Multimedia Content from a Streaming Media Server. <http://www.leeds.ac.uk/iss/digitaldelivery/>
8. JPEG 2000 The new standard. <http://www.jpeg.org/jpeg2000/>
9. DROID (Digital Record Object Identification). <http://droid.sourceforge.net/wiki/index.php/Introduction>
10. National Library of New Zealand Metadata Extraction Tool Version 1.0 <http://www.natlib.govt.nz/en/whatsnew/4initiatives.html#extraction>
11. Proposal for a minimum preservation metadata element set based upon the PREMIS data dictionary.

## MIDESS: WP 5 – Digital preservation requirements specification

[http://ahds.ac.uk/about/projects/hybrid-archives/wp44\\_preservation\\_metadata.pdf](http://ahds.ac.uk/about/projects/hybrid-archives/wp44_preservation_metadata.pdf)

12. Digital Images Archiving Study. Arts and Humanities Data Service (AHDS).  
([http://www.jisc.ac.uk/uploaded\\_documents/FinaldraftImagesArchivingStudy.pdf#search=%22FinaldraftImagesArchivingStudy.pdf%22](http://www.jisc.ac.uk/uploaded_documents/FinaldraftImagesArchivingStudy.pdf#search=%22FinaldraftImagesArchivingStudy.pdf%22))

13. PREMIS. Preservation Metadata Maintenance Activity.  
<http://www.loc.gov/standards/premis/>

14. National Information Standards Organization technical metadata for still images.  
[http://www.niso.org/standards/standard\\_detail.cfm?std\\_id=731](http://www.niso.org/standards/standard_detail.cfm?std_id=731)

15. MIX NISO Metadata for Images in XML Schema. <http://www.loc.gov/standards/mix/>

16. METS. Metadata Encoding and Transmission Standard.  
<http://www.loc.gov/standards/mets/>.

17. WC3 Date and Time formats. [www.w3.org/TR/NOTE-datetime](http://www.w3.org/TR/NOTE-datetime)

18. Moving Images and Sound Archive Study.  
[http://www.jisc.ac.uk/uploaded\\_documents/Moving%20Images%20and%20Sound%20Archiving%20Study1.doc](http://www.jisc.ac.uk/uploaded_documents/Moving%20Images%20and%20Sound%20Archiving%20Study1.doc)

19. Wikipedia. Pulse Code Modulation. [http://en.wikipedia.org/wiki/Pulse-code\\_modulation](http://en.wikipedia.org/wiki/Pulse-code_modulation)

20. Wikipedia. Advanced Audio Coding.  
[http://en.wikipedia.org/wiki/Advanced\\_Audio\\_Coding](http://en.wikipedia.org/wiki/Advanced_Audio_Coding)

21. Wikipedia. MP3. <http://en.wikipedia.org/wiki/MP3>

22. Wikipedia. Broadcast Wave format. <http://en.wikipedia.org/wiki/BWF>

23. Wikipedia. Audio Interchange File format. <http://en.wikipedia.org/wiki/Aiff>

24. National Library of New Zealand.  
<http://www.natlib.govt.nz/bin/media/pr?item=1154899822>

25. National Digital Heritage Archive Programme / Project Kronos.  
<http://www.natlib.govt.nz/files/initiatives/NDHA%20Kronos%20Partnership%20Q&As%20Aug%202006.DOC>

26. Dspace Digital Repository Software. <http://www.dcc.ac.uk/resource/technology-watch/dspace/>

27. Digital Preservation. Architecture and Technology for Trusted Digital Repositories. D-Lib Magazine June 2005 Volume 11 Number 6. <http://www.dlib.org/dlib/june05/jantz/06jantz.html>