DIGITAL AV MEDIA DAMAGE
PREVENTION AND REPAIR

Initial IT-based strategies for avoiding, mitigating and recovering from digital AV loss
&
Initial conceptual risk management framework and tools for digital AV preservation

Deliverable D3.1

SEVENTH FRAMEWORK PROGRAMME

DAVID identifier:	DAVID-D3.1-ITInnov-Initial-Strategies-and-Risk-Framework-FINAL.docx
Deliverable number:	D3.1
Author(s) and company:	M. Hall-May (ITInnov), G. Veres (ITInnov), J-H. Chenot (INA), C. Bauer (ORF), W. Bailer (JRS)
Abstract:	Digitised and born-digital AV content presents new challenges for preservation and long-term quality assurance. Archives have rapidly developed strategies for avoiding, mitigating and recovering from digital AV loss using IT-based systems. A risk-based framework based on the essential properties of AV assets and documented preservation metadata is required to determine how best to minimise the risk of digital damage.
Internal reviewers:	J-H. Chenot (INA) C. Bauer (ORF), R. Meszmer (ORF)
Work package / task:	WP3 T3.1 & T3.2
Document status:	Final
Confidentiality:	Public

Version	Date	Reason of change
1	2013-10-14	Document created (e.g. structure proposed, initial input…)
2	2013-11-22	Initial contributions integrated
3	2013-11-25	Executive summary, conclusions, glossary and references
FINAL	2013-11-28	Final version addressing QA comments

Acknowledgement: The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 600827.

Disclaimer: This document reflects only the author's views and the European Union is not liable for any use that may be made of the information contained therein.

This document contains material, which is the copyright of certain DAVID consortium parties, and may not be reproduced or copied without permission. All DAVID consortium parties have agreed to full publication of this document. The commercial use of any information contained in this document may require a license from the proprietor of that information.

Neither the DAVID consortium as a whole, nor a certain party of the DAVID consortium warrant that the information contained in this document is capable of use, nor that use of the information is free from risk, and does not accept any liability for loss or damage suffered by any person using this information.

1 Table of Contents

1 Table of Contents ... iii

2 List of Figures ... v

3 List of Tables ... vi

4 Executive Summary ... 7

5 Introduction ... 9
5.1 Purpose of this Document ... 9
5.2 Scope of this Document ... 9
5.3 Status of this Document ... 9
5.4 Related Documents ... 9

6 Problems affecting Digital AV Contents ... 10
6.1 Digital versus Analogue Audio-visual Content ... 10
6.2 What is Digital Damage? ... 10
6.3 Preservation Metadata for Digital AV Content ... 14

7 IT-based strategies against Digital AV Loss ... 16
7.1 Avoiding, mitigating and recovering from digital AV loss ... 16
7.1.1 Robustness of Technology ... 16
7.1.2 Robustness of People ... 22
7.1.3 Robustness of Processes ... 23
7.2 IT-based preservation at INA ... 24
7.2.1 Physical security of AV data ... 24
7.2.2 Backups of AV data ... 25
7.2.3 System security of AV data ... 25
7.2.4 AV data Security overall in INA ... 25
7.2.5 Comparison with analogue workflows ... 25
7.3 IT-based preservation at ORF ... 26

8 Conceptual Risk Management Framework ... 28
8.1 Archive risk management ... 28
8.1.1 Risk management in safety-critical engineering ... 28
8.1.2 Financial risk management ... 29
8.1.3 Archive risk management ... 29
8.1.4 Workflow/business process risk modelling ... 30
8.2 Risk Management Model ... 31
8.2.1 Risk management steps ... 31
8.2.2 Risk measures ... 32
8.2.3 Classification of threats/risks in digital preservation ... 34
8.3 Analysis of ORF risk management ... 35
8.3.1 Identifying workflows and processes taken place during digital preservation at ORF ... 35
8.3.2 Defining objectives of preservation process and risk management ... 37
8.3.3 Identifying risks/threats, negative consequences, controls with cost and time and their analysis ... 37
8.3.4 Classification of risks/threats according to SPOT model ... 38
8.3.5 Further work on Risk management of ORF DiMi workflow ... 39
8.4 Initial Tools for Risk Management ... 39
8.4.1 Background ... 39

© DAVID consortium: all rights reserved

8.4.2 BPMN 2.0 ..... *40*
8.4.3 Archive Model *40*
8.4.4 Risk Model *41*
8.4.5 Model Class Diagrams *41*
8.4.6 Simulation Model *42*
8.4.7 Further Work to be done *43*
8.5 Preservation metadata model 43
8.5.1 Scope *44*
8.5.2 Model definition *44*
8.5.3 Relation to BPMN *46*
8.5.4 Relation to other preservation data models *46*
8.5.5 Specific activities *47*
8.5.6 Properties *48*
8.5.7 Tools *48*

9 Conclusions 49

10 References 50

11 Glossary 55

© DAVID consortium: all rights reserved

2 List of Figures

Figure 1: Corruption of artefact-free video during MPEG-2 to DVD conversion 13
Figure 2: Risk functions: graphical representation of VaR, VaR deviation, CVaR, CVaR deviation, Maximum Loss and Maximum Loss Deviation 33
Figure 3: DiMi Workflow model 36
Figure 4: Archive Risk model 41
Figure 5: Archive Risk simulation model 42
Figure 6: Entities of the preservation data model, their relations and the most important core properties. Blue entities are related BPMN entities. 44

© DAVID consortium: all rights reserved

3 List of Tables

Table 1: Example classification of digital damage 13
Table 2: Identification of risks/threats in DiMi workflow - extract 37
Table 3: Classification of risks/threats according to SPOT model 39
Table 4: Description of the model entities. 45
Table 5: Additional properties for specific tool types. 48

© DAVID consortium: all rights reserved

4 Executive Summary

Preservation aims to ensure that cultural heritage is accessible for the long term. From the 20th century onwards, audio-visual (AV) content has provided a significant record of cultural heritage. Increasing volumes of AV content that have been digitised from analogue sources or produced digitally present new preservation challenges. The focus is no longer on reducing damage to the physical carrier by maintaining a suitable environment; rather, archives must ensure that the significant characteristics of the content, represented digitally, are not lost over time. Digital data enables easier transfer, copying, processing and manipulation of AV content, which is at once a boon but also a problem that requires continuous and active management of the data.

Digital damage is defined for the DAVID project as any degradation of the value of the AV content with respect to its intended use by a designated community that arises from the process of ingesting, storing, migrating, transferring or accessing the content. In order to understand where in the life cycle of a digital object damage has occurred, it is important to understand the provenance and history of the artefact. This leads to a requirement for a record of preservation metadata.

Archival processes dealing with digital AV are underpinned by IT systems. In the few years that archives have been working with digitised and born-digital content, best practice in terms of digital contents management has rapidly evolved. Strategies for avoiding, reducing and recovering from digital damage have been developed and focus on improving the robustness of technology, people and processes. These include strategies to maintain integrity, improve format resilience and interoperability, and to combat obsolescence. Redundancy and simplification (of technology and processes) are critical, as well as quality checking, change management and audit. Both the French national archive, INA, and the Austrian broadcaster ORF, deal with digital AV content and use a mixture of such strategies to minimise the risk of damage occurring.

Understanding where and how risks arise and which strategies to use to treat them is part of risk management. Safety-critical systems engineering (in which failure results in injury or loss of life) and financial systems management (in which consequences are economic in nature) have both produced tools and techniques for modelling and assessing risk. In the archive domain, the OAIS standard requires preservation planning to produce periodic risk analysis documents but does not mandate a process. Several research and development tools have been produced to aid in preservation planning, mostly focusing on the analysis of the cost of preservation versus the risk of loss of AV assets.

In the burgeoning domain of business process modelling (and execution), it has recently been noted that risk management can be combined with workflow specification. Many techniques are based on the BPMN standard. Nascent techniques exist for static analysis of risk in a business process, such that risk mitigation strategies can be optimised. There is also early research into capturing information from executing business processes for real-time risk analysis.

A conceptual risk management model has been developed in DAVID, which identifies risks to steps of a workflow and classifies their impact according to a threat model (SPOT) that focuses on the essential properties of the AV asset (including identity, persistence, and renderability). Controls that can be used to mitigate the risk can be identified, and the approach uses the concepts of expected loss and (conditional) value at risk to measure risk and to inform the placement and effectiveness of controls in digital preservation workflows. ORF's Digital Migration (DiMi) workflow has been modelled and analysed using the proposed approach. Risk information was captured from experts within ORF, who classified the risks arising in the workflow steps according to their likelihood and severity using the SPOT model.

An early prototype tool has been developed that extends the Oryx business process modelling framework in order to tailor it to preservation workflows and to capture risk-based information (such as failure probabilities and control effectiveness). The tool supports the proposed risk model and allows the user to input results of an inductive risk analysis. The intention is to use the models produced using this

© DAVID consortium: all rights reserved

tool in a simulation-based analysis that will allow a decision-maker to deduce where in the process failures are likely to cause unacceptable digital damage and how to use risk treatment strategies to prevent, reduce or recover from such damage.

A preservation metadata model has been developed in the project, which will allow capture of the historic creation and processing of AV assets in a common structured format. This is useful for manual analysis and audit but will also prove a useful input to the risk model. Current analyses are based on expert judgement and limited historical data on failure, as failures are usually fixed and forgotten during times of crisis, and are often not rigorously documented. It is anticipated that the preservation metadata will allow us to discern which information to capture in order to feed back to the simulation-based analysis approach and improve the output's relevance for decision-makers.

© DAVID consortium: all rights reserved

5 Introduction

5.1 Purpose of this Document

The purpose of this document is to motivate and describe work carried out towards providing a risk-based framework for managing the long-term quality assurance of digital AV content. This combines the understanding of the ways that loss can occur (from DAVID project WP2) with the strategies that can be used to minimise the risk of loss, given the constraint of finite resources. The report includes work carried out in tasks T3.1 and T3.2 of work package 3. The task descriptions are as follows:

T3.1 How can the loss of digital AV content be prevented, mitigated and recovered? This task aims at defining the strategies that can be used to reduce the probability of loss, reduce the impact of loss and recover from loss events for different content types and preservation systems through the use of IT-based technologies.

T3.2 How can the archive minimise the risk of loss of content (that results in an unacceptable degradation in the usability of that content) and assure quality in the long term? This task aims at defining a conceptual risk framework that allows the archive to assure the long-term usability of digital AV content (i.e. to minimise the risk of loss within the bounds of the available resources required to maintain a given level of usability).

5.2 Scope of this Document

This report covers work done in the first year of the DAVID project under tasks T3.1 and T3.2, including conclusions drawn from the work done in WP2, tasks T2.1 and T2.3. It does not include work carried out in task T3.3 - *Recommendations and techniques for creating new content in a 'born robust' form*.

5.3 Status of this Document

This is the final version of the document.

5.4 Related Documents

It is recommended that the reader familiarise themselves with the contents of deliverable D2.1 - Damage assessment data and infrastructures, assessing consequences of loss on usability, available at http://david-preservation.eu/wp-content/uploads/2013/10/DAVID-D2-1-INA-WP2-DamageAssessment_v1-20.pdf.

© DAVID consortium: all rights reserved

6 Problems affecting Digital AV Contents

6.1 Digital versus Analogue Audio-visual Content

Until recently, the preservation of analogue content has been intrinsically linked to its method of production; specifically, the media that is used to carry the signal (the carrier). This means that archives preserved 'masters' on magnetic tape, film and even phonograph cylinders [CPDP]. Where masters no longer exist or content was not professionally produced, archives have been forced to preserve 'access' copies on media such as vinyl records, VHS/Betamax tapes, and audio cassettes. To reduce the risk of damage, archives had to consider the physical characteristics of the media and care for the physical environment to which the media was sensitive (e.g. light, heat, humidity, dust) and to look after the machines that read the media. To increase the chances of being able to read the content again, archives often created copies of the artefact, in case one copy was damaged.

While analogue replication is possible, such replication is inevitably imperfect and some might argue that part of curation has traditionally been to maintain the 'original' copy, as some of the value of the asset is as much in the 'carrier' as in the 'essence'. In some communities, the preservation of a physical work of art, such as an oil painting, is still the principal objective, as the value of the artefact is in its uniqueness (its look and feel, the brushstrokes). However, even within these communities, it has been recognised that digital imaging can produce an artefact in which ageing has been arrested and which can (potentially) outlive the original.

Digital content (digitised or born digital) can be copied, transferred, shared and manipulated far more readily than its analogue equivalent. This presents us with a different preservation challenge and one with which we are just getting to grips. As archives have started digitising their existing content and producers begin to submit born-digital content, the challenge is less and less about preservation of the physical artefact — i.e. the audio tape, the film reel, the wax cylinder — and far more about maintaining bit-perfect copies in the short-term and ensuring that the migration to new formats preserves significant characteristics of the content in the long-term.

In a world of digital AV content, preservation is largely agnostic to the carrier that is used to store and deliver the content. Therefore, preservation and archiving is about making sure that the digital data is *safe* and that processes that manipulate the data do not cause *damage*. There are many strategies, tools and techniques for avoiding, reducing and recovering from digital damage. These will be investigated in section 7, but first we must define what is meant by 'digital damage'.

6.2 What is Digital Damage?

It is critical to define the scope of the term 'digital damage' for the DAVID project and for the techniques, tools and recommendations that it will develop.

Digital damage is any degradation of the value of the AV content with respect to its intended use by a designated community that arises from the process of ingesting, storing, migrating, tranferring or accessing the content.

The above definition may seem broad. Indeed, it covers damage arising from failure of the equipment used to store and process digital content, as well as that arising from human error or from 'failure' of the process. The definition is focused on preservation processes, but is not limited to damage at the data or bit level; rather, it encompasses damage to the content, metadata and structure of the asset. Damage arising from natural disasters, such as the recent destruction of digital artworks by Hurricane Sandy [Chayka, 2012], is excluded.

The focus of the DAVID project is on preservation processes underpinned by information technology (IT). As such, the definition of digital damage does not make provision for loss arising through, for

© DAVID consortium: all rights reserved

example, preservation selection policy or other 'out of band' administrative decisions or changes. For example, the definition does not cover a change in rights, which most certainly would degrade the value of an asset if the community no longer has the right to use the content as it wants. On the other hand, if the rights of an AV asset could not be ascertained, as they had been lost during ingest, storage, migration, transfer or access, then this would be considered a case of digital damage.

Analysis of current samples of digital AV content in DAVID project WP2 work package provides a vocabulary for talking about damage. Threats to data integrity are a well understood problem in digital preservation and has mature solutions, which will be reviewed in Section 7. In terms of AV content, digital damage presents itself through a number of artefacts, such as dropouts and noise, which require quality checking (QC) processes to detect and repair; these are the remit of WP4. Thornier problems are those that arise from digital file format and tool incompatibility, as well as capturing preservation actions as metadata for audit purposes. These latter problems are addressed in WP3.

Damage must be related to future use, designated communities and the content's significant characteristics; content that is considered 'damaged' by one person might be acceptable to another. It is difficult to predict the future use of content (and the future tools that will process it) and so ensure that it will be acceptable for use (i.e. considered 'undamaged'). The default position for archives thus far has therefore been by necessity to 'throw away as little as possible'. Content use that must be considered is reuse (e.g. re-editing into a new programme), access (both now and in the future through new channels) and regulatory compliance (in which what is considered 'good enough' to comply with legal requirements may change).

Unfortunately, resources are finite and so archives must (increasingly) make preservation trade-offs with respect to time and capacity requirements (and hence cost). In production it is typical to discard 'rushes', different camera feeds and to keep (for the long term) the edited version only. However, this is an accepted process from the days of producing analogue content; hence, the argument is that digital workflows that discard such information are 'no worse' than their analogue predecessors.

With the advent of digitisation, many processes must inevitably sample, quantise and throw away information (such as non-visible light), which could be used in the future when digitisation or playout techniques have improved. Similarly, born-digital content often suffers from resampling, colour-space changes and lossy compression algorithms, which have been built on the premise that the information 'thrown away' is not perceptible by human senses, but this precludes future uses in which the beholder is a machine [Sitts, 2000], which could make use of such 'hidden' information.

There are public efforts to collect samples of damage, such as the community-supported A/V Artifact Atlas [AVArtifactAtlas] and the Flickr Atlas of Digital Damages [LeFurgy, 2012]. In some cases, the crowd-sourced effort includes recipes for recovering from the damage, such as the Open Planets Foundation wiki [Wheatley, 2013]. This wiki categorises digital preservation and curation issues using the following broad classification (abbreviated), some of which include digital damage under the definition proposed above:

- Appraisal issues
- Conformance issues
- Bit rot issues
- Contextual issues
- External dependency issues
- Quality issues
- Obsolescence issues
- Rights issues

Damage arising from corruption during storage, whether it be latent damage or caused during a failed read/write of data, has been the focus since large-scale digitisation of AV assets has been undertaken by major archives. It was one of the focuses of the FP7 PrestoPRIME project [PrestoPRIME], which

© DAVID consortium: all rights reserved

produced best practice guidelines and tools that have had significant influence within the AV preservation community.

Corruption during storage can cause the following damage:

- Corruption of essence causing visible/audible artefacts during playback, e.g. blocks, dropouts, out-of-sync audio, stops or stutters, or preventing migration to another format.
- Corruption of wrapper preventing playback, causing degraded playback or preventing automatic migration to another format.
- Corruption of metadata affecting identity of the asset (e.g. for search), preventing playback or causing degraded playback and preventing automatic migration to another format.

The above problems are significant, but analyses carried out in DAVID project WP2 revealed that they rarely occur. If such corruption does affect an asset, restoring the asset from a replica or, in the worst case, from the original source, is a tried-and-tested solution. According to these analyses, much greater problems arise from the following processing of the digital AV assets that do not necessarily result from corruption:

- Encoding of essence that renders it unplayable, causes artefacts on playback, or (undetectably) causes problems for later playback or migration (using new or upgraded tools).
- Encoding of wrapper that introduces damage with similar consequences as above, e.g. as a result of missing, unexpected, incorrect or ambiguous encoding of data such as field dominance, time codes, closed captions, additional audio tracks.
- Encoding of metadata that that affects the asset identity or causes damage with similar consequences as above, e.g. as a results of missing, unexpected incorrect or ambiguous metadata.

These problem types are potentially much more damaging, as they result from systematic failure, as opposed to the random failure of corruption, and so could affect whole batches of assets. Ensuring interoperability of files within the current operating environment as well as with future changes to this environment is a critical consideration when generating digital assets.

Encoding that causes problems that are not immediately apparent (e.g. during QC or playback), but which causes migration to fail or produces a next-gen file with artefacts are particularly insidious. For example, Figure 1 shows the results of converting an MPEG-2 file, which plays without artefacts on a number of common video players, to DVD [VideoHelp, 2008]. Owing to different interpretations of AV standards by software programmers such errors will arise in the implementation of different player and migration tools. Many standards include ‘reserved’ elements of the file for later extension of the standard, which should be ignored by current software implementations. If current tools write data in these reserved areas, this might be interpreted (wrongly) by future implementations, causing problems affecting all files processed using this tool.

© DAVID consortium: all rights reserved

Figure 1: Corruption of artefact-free video during MPEG-2 to DVD conversion

Further documented examples of damage caused by encoding failures are where the profile used to create the file is unknown (i.e. not recorded in the file or in metadata) [Wheatley, 2011], where the file (or frame) is truncated [Chivers, 2012a], and where subtle artefacts are introduced during encoding [Chivers, 2012b].

One solution to the problem of damage caused by context change is to control the change to the operational environment. Preservation workflows are often complicated, involving many different tools. Firmware and equipment are changed frequently (with respect to the lifetime of the AV asset) and can have devastating effects on content access. Upgrading a part of the workflow (e.g. introducing a new tape reader or even new firmware on an existing device) might mean that the content cannot be reliably read back or interpreted, or that it cannot be written out in a way that can be interpreted in the future. It is important to focus on the critical points in workflows at which changes (detectable or otherwise) can have an effect on content usability.

Using WP2 analyses, it is possible to begin a risk assessment by classifying the issues according to:

- the process in which the problem is detected
- the nature of the problem (i.e. immediately detectable effects)
- the cause of the problem

Table 1: Example classification of digital damage

Process	Problem	Cause
Access/playout	Content cannot be played	Format incompatible with playout system owing to file ingest/migration problems
		Format incompatible with playout system wing to operating environment change
		Playout malfunction
	Content displays artefacts during playout	Artefacts introduced from corruption during storage
		Artefacts introduced during ingest/migration
		Artefacts existed in source prior to ingest
		Playout malfunction

© DAVID consortium: all rights reserved

Ingest	Format not identifiable	Format identifier not present in submission
		Format/profile is unknown
	Metadata cannot be ingested	Metadata not present in submission
		Metadata in unknown/incompatible format
		Metadata incomplete
Storage/scrubbing	Integrity check fails on scrubbing	Checksum miscalculation
		Checksum missing
		Latent corruption since last integrity check
		Corruption on write
		Corruption on read
Migration	File unreadable by migration tool	Storage read error
		File not at specified location
	Migration fails	Corruption on read
		Migration tool does not support source format
		Migration tool does not support destination format

From the above, it should be clear that there are two distinct aspects of digital 'damage':

- corruption during storage or transfer of an existing asset (i.e. the data changes with respect to a reference)
- damage introduced during creation or processing of a new asset (i.e. there is no reference)

Good strategies and best practice exist to deal with the first kind of damage. The second aspect of digital damage above affects the ability to use the content, where 'use' primarily means to open and 'render' the content using tools available at the time of access [Cochran, 2012]. However, preventing such damage requires understanding the workflow in which assets are created and processed. To help with risk assessment of workflows, it is necessary to be able to capture the history of a file, so that we can determine at which point in the workflow damage occurred. This requires a structured form to be able to capture as metadata the processes that have been executed on a digital object.

6.3 Preservation Metadata for Digital AV Content

This section lists preservation metadata that could be used in addition to basic technical metadata (structural metadata) in order to mitigate the risks related to the problems described in D2.1. It can also serve as a basis for defining metadata for born-robust content.

Well-defined specification of wrapper and codecs. While technical metadata models typically include some information about formats, it is important to unambiguously specify wrapper and codec formats and their variants (profiles, versions, operational patterns, etc.). This should be done using a controlled vocabulary. As this information continues to grow as new technologies are developed, it should preferably be maintained in a registry managed by a trusted organisation.

Structural relations. It is necessary to document relations between the resources constituting a complete representation of the digital item, including all the alternative representation included in the preservation package (which might share resources).

Fine-grained checksums. For detection of corruption and repair, it helps to have checksums on fine temporal granularity (e.g. frame, GOP).

Cross-check QA metadata. Quality metadata on cross-check between metadata and actual content properties is needed to detect inconsistencies and avoid incorrect processing.

© DAVID consortium: all rights reserved

Embedded objects. If objects are embedded in containers (wrappers, packages), detailed information on embedded objects and their type, encoding, language, etc. is needed.

External information. If possible, any dependencies to information outside the preservation package should be avoided. If needed, the references to external information need to be well documented, including their type and sufficient identification of the external resources.

Playback environment. Playback environments for specific content types should be well documented. As this information is dynamic, it should not be kept in the item metadata, but in a separate database. Given comprehensive documentation of content properties, matching playback environments at the time and place of access can then be found. As the information about playback environments may be changing often, and it may be hard for a single institution to keep track, it should preferably be maintained in a registry managed by a trusted organisation.

Rights. Comprehensive, fine-grained and machine-processable rights metadata should be kept with the content, using e.g. recently proposed standards such as MPEG-21 CEL/MCO [CEL, MCO].

Process metadata. All processes that lead to current version of the item should be documented, including the involved tools and their parameters.

© DAVID consortium: all rights reserved

7 IT-based strategies against Digital AV Loss

7.1 Avoiding, mitigating and recovering from digital AV loss

When occurring, digital damage can cause some level of loss in an audio-visual asset. There is always a risk of loss to digital content. The risk embodies the likelihood that a threat to the content will occur, causing loss, and the impact of the loss. There are essentially three approaches for treating such risks, as follows:

Avoid Through definition of preservation processes and/or the choice of tools, the aim is to make it nearly impossible for loss (of a particular kind) to occur. This strategy is relevant at the planning or replanning stage and essentially embodies a 'find another way' approach. Of course, finding another way inherently involves trade-offs, as the alternative approach may not meet the same requirements as the original. As an extreme example, it is possible to avoid digital loss entirely by deciding against digitisation, but this raises other risks associated with the on-going preservation of analogue material. A more realistic strategy is to avoid the risks inherent in migrating lossy encodings by choosing only lossless formats for archive.

Mitigate Risk mitigation strategies aim to reduce the likelihood that the risk will occur and/or reduce the impact if it should occur. In preservation of digital assets, many strategies fall into this category and range from the choice of equipment, formats and definition of processes, such that they are robust to failure.

Recover Recovery is a kind of strategy that relies on detection of an undesirable state (i.e. the risk has occurred) and (perhaps imperfect) transition to a better state. To be effective, recovery relies on a level of redundancy in the system. The tasks of detection and recovery can be a manual, automatic or hybrid solution. Where the choice of technology incorporates failure detection and recovery at a low level (e.g. bit-level corruption detection and repair on read using CRC codes), this can seem like a mitigation strategy (e.g. the storage technology is robust to failure through error concealment). If redundant information does not exist, allowing the system to recover the loss perfectly, imperfect recovery can often be achieved through interpolation (e.g. partial repair of a master file through inter-frame interpolation or frame duplication).

A fourth category, transfer of risk, in which the financial impact of the risk is borne by another party, is not considered here.

The information technology that underpins the processing and storage of digital AV content is all important to its preservation. Digital damage can arise through failure of the technology, failure of the operator to use the technology correctly, or through inappropriate use of technology (or a combination of technologies) as part of a larger process. Therefore, in an IT-based digital archive, there are three areas in which the above types of risk treatment strategies can be employed: robustness of technology, of people and of processes.

7.1.1 Robustness of Technology

Commensurate to the rise in digital AV content in today's archives is the use of information technology to manage and store this content. The choice, maintenance and composition of technology can have serious implications for the risk to which AV assets are exposed. The archive can often exercise control over the choice and use of hardware equipment, software tools, and digital formats. The following strategies are relevant for technology-related risks.

Integrity

Archives are consumers of technology and do not tend to build storage and processing technologies from scratch; rather, they compose appropriate technological components to create a preservation

© DAVID consortium: all rights reserved

system. Therefore, choosing technologies that purport to be reliable is a good starting point for building a reliable preservation system that maintains data integrity.

Storage devices, such as spinning hard disk drives, solid-state drives, and LTO tapes, as well as the devices required to read them, such as servers and tape robots, are vulnerable to a number of threats. Mechanical failures can occur through wear and tear (e.g. tape or cassette breakage, HDD head crash) or mishandling. Electrical failure (e.g. power surge, outage or spikes, electromagnetic interference) can cause permanent or transitory damage to data. Firmware bugs, such as in the device control software, can lead to many kinds of errors, including miswriting data to the wrong location.

Failures can occur when reading from or writing to a storage device. Such failures are *active* failures and may have one of the causes listed above, i.e. mechanical, electrical, or software (e.g.firmware) bug. To improve the chances of maintaining data integrity against such threats, archives typically choose quality components, in which both the software and hardware have been rigorously tested by the vendor. In addition, the archive must perform regular monitoring and maintenance on the devices to detect (and ideally predict) failure. Older devices that are used to store data should be 'refreshed' after a certain period of time, whereupon the data is copied to a new storage device and it is verified that the transfer was successful and correct.

Failures can also occur during storage, while the bits remain untouched on the device. These failures are *latent* failures and are sometimes referred to as 'bit rot', in which single bits change their value. The frequency of these 'bit flips' depends on the way in which the particular carrier represents binary data on its media, and on the physical density of the data.

The failure mode of a storage device means that some amount of data that is read from the device is either not accessible or is not the same as that which was previously written to the same location. The amount of data that is corrupted or inaccessible can be a single bit, bye, block, sector or the entire contents of the device in the case of wholesale unit failure. Such errors can occur silently, in that they are not detected at the time of corruption. To protect against such errors, manufacturers build error checking codes (such as CRCs or Oracle's T10-DIF feature [Petersen, 2007]) into the data stored by the device. This is used to check the values of the data when read from the device and can either detect or, in some cases, correct instances of corruption. These codes use the strategy of *redundancy* (see below).

Even using error correcting codes, there is a chance that some errors cannot be corrected and must be dealt with at higher levels of the preservation system 'stack'. Typical values that are cited by manufacturers for unrecoverable bit error rate (BER) are 1 in 10^{14} for HDD and 1 in 10^{17} for LTO tape. However, CERN's analysis [Panzer-Steindel, 2007] of storage devices observed an actual BER of around 3×10^{7}, given the combined failure rates of all the devices used in the end-to-end chain of data movement operations (e.g. CPU cache, system memory, disk controller, network card).

Given the likelihood of silent data corruption occurring at the storage layer, even when using reliable components, typical strategies rely on 'defence in depth'. The layers above the storage layer — the file system, the operating system, and the application — can also help to detect and correct errors to maintain data integrity. Using a file system, such as ZFS or IRON FS [Prabhakaran, 2005], that assumes the underlying storage devices to be unreliable, provides extra protection for data integrity. However, even these file systems can exhibit failure modes [Sun, 2010], and so the above layers must detect and correct them.

To ensure data integrity, each read or write operation must be verified to have been performed correctly. Verification of each step creates a 'chain of custody', which should start as early in the digital object's lifetime, ideally at its creation.

As storage devices are fallible, error checking and media refresh are inevitable. However, recent developments in storage technology promise long-lasting, error-free digital media [Zhang, 2013], [de

© DAVID consortium: all rights reserved

Vries, 2013]. The amount of trust to put in such technologies, when they become available, is still open to debate, but it is clear that even with completely reliable digital storage, we must still concern ourselves with what the digital data represents, i.e. the 'format' of the data. In the end, errors resulting from 'chattery' cables, dodgy disc controllers or flaky firmware (and human error higher up the stack) are much more likely than corruption resulting from cosmic rays.

Format resilience

Assuming that the underlying software and hardware stack may be unreliable and therefore, on occasion, some amount of data may be corrupted, one strategy to reduce the impact of this corruption is to choose a file format that is resilient to this corruption.

The CERN study, mentioned above, noted that when considering compressed files (e.g. zip archives) a single bit error causes the whole file to be unreadable (with a 99.8% accuracy). The choice of file format has a significant impact on the overall loss rate, even if high integrity components with low bit error rates are used.

While corruption in compressed files often leads to greater impact on the AV contents than in uncompressed files [Gattuso, 2013], the kinds of compression currently used in AV formats are such that the effects of corruption are often limited in extent. Whereas the failure modes of archive compression standards, such as zip, gzip, arc, tend to render the whole file unreadable, AV formats often use inter-frame or intra-frame encoding, which limits the effects of corruption to a section of the file (e.g. an MPEG-2 GOP, or one image). However, it is worth considering a file format that is resilient to the kinds of failure modes that the storage layer exhibits. Heydegger proposes using bit error resilience to determine the robustness of a file format [Heydegger, 2008] and analysed several image formats for the effects of corruption [Heydegger, 2009].

Existing AV codecs are typically optimised to recover from (or conceal) errors in transmission, but the failure modes of storage systems are different to those of transmission. Transmission codecs are optimised for the error characteristics of a communication channel, e.g. random errors in wireless communications, data loss due to channel congestion. File codecs must be resilient to the failure modes of storage: latent corruption (bit rot), human error leading to carrier damage, read/write errors due to device driver bugs. Owing to the large volumes of data handled within an archive, transfer between systems is typically over local fast and reliable networks, as opposed to the unreliable transport mechanisms (e.g. digital television broadcast) typically used to transmit content to viewers.

In the still image preservation domain, there is concern over the suitability of formats (such as JPEG2000) to fulfil simultaneously access and preservation requirements [LeFurgy, 2013]. Buonora and Liberati conclude that JPEG2000's error control features against failures in transmission give it an advantage over other formats for preservation [Buonara, 2008]; however, their conclusions do not extend to the use of JPEG2000 as an essence encoding format in the AV domain. INA's analysis of next-generation master formats [Varra, 2012] concluded that lossless JPEG2000 (in an MXF wrapper) [MXF_OP1a_JP2_LL] is a suitable format for SD content, while 'visually lossless' JPEG2000 (lossy encoding at 200Mbit/s) [MXF_OP1a_JP2_LSY] is suitable for HD content, given the trade-off between storage and quality requirements. The key criteria considered in the analysis were interoperability (compatibility) and quality, while robustness to corruption was not considered.

The BBC [Weerakkody] proposed an archive 'format', which reorders and replicates the bits so that typical failure modes of the carrier have less impact on the high-value content (e.g. header/metadata, low-frequency data). Such approaches have inevitable trade-offs, in that the format may not necessarily be optimised for immediate playback (e.g. for broadcast), because playing the first chunk might involve reading the whole file (in case the audio is at the end of the file). As such, the file might have to be transformed into a playable format; however, this is the case for uncompressed AV, so the overhead is already accepted by some archives/broadcasters.

© DAVID consortium: all rights reserved

MXF considers the relationship between the layout of the file structure and the underlying storage medium: the KLV alignment grid (KAG) allows padding to be inserted, such that 'important' parts of the file are aligned with the sectors of the storage device, thereby improving performance, but more importantly allowing us to minimise the areas of the file that are corrupted if a particular sector fails.

The ideal format would offer 'graceful degradation', so that as the amount of corruption increases, the quality of the content decreases gradually, rather than failing utterly. In this sense, uncompressed is better than compressed formats, but increases the file size considerably. Similarly, wavelet-based compression offers an arguably better degradation profile than DCT-based compression, as the former spreads the effects of the error over the whole image, which may be difficult to notice in one frame of a file, while corruption of the latter affects a single block of the image, which stands out in a sequence of images.

The choice of format does not only affect robustness of a single file to corruption. We must also consider the relationship to other files, in which loss of an external reference (e.g. separately stored audio tracks) severely degrades the AV asset. A single 'format' often offers different profiles, some of which pack all information into one file (e.g. MXF OP-1a), while others split the content (video, audio, metadata) into several files (e.g. MXF OP-Atom). An archive-specific variant of MXF, called MXF AS-07, is currently being developed.

PHENICS project has recorded live musical performances using multiple camera angles and EEG (gesture) capture and has created a repository of these works, combined with the score. Each 'data pack' represents one performance and contains several video files, CSV files for gesture information, and XML files for other metadata, which are all related using entries in a SQL database. This is clearly a rich set of data, the loss of any part of which would diminish the value of the recording.

Given that some corruption may occur, the choice of file format also affects the ease with which the content can be repaired. For example, in a format that uses a table of offsets in the header to indicate the relative position of frames in the file, corruption of the header can cause the rest of the contents to unreadable (or at least unseekable), as it is not clear where each frame begins. This offset table can be recovered by inspection of the frame data, but the process is unlikely to be simple. If the boundaries of the frame data are marked, then this process becomes easier, but typically offset tables are used for rapid seeking to a particular point in the video. As a concrete example, recovery is made easier if we were to use Constant Bytes per Element (CBE) essence, such as IMX, rather than Variable Bytes per Element (VBE), which includes MPEG long GOP, because each CBE frame is of the same size. We can then reconstruct the index table from only a single frame.

No matter how good the format, the interpretation of the format is reliant on the implementation of the codec and of the application using that codec. Rosenthal observed that Postel's law should apply when conforming to standards: be strict in what you emit, liberal in what you consume [Rosenthal, 2009a]. This maxim should lead to improved robustness and is one way to solve the insoluble problem of (inevitably buggy) software generating malformed output that is in some way incompatible with other software. The tools used to access the content (e.g. for playout or migration) must interpret the standard/format as loosely as possible (and couple this with QA of the output/migrated content). Fixing the bug is not a solution, because the already generated content is still wrong and is expensive (or impossible) to repair. In Rosenthal's example, he noted that web crawlers do not reject W3C non-compliant HTML.

Interoperability

Even seemingly well-formed files, which have been verified to be free of corruption, can also cause problems. Incompatibility of codecs/wrappers (or specific implementations of these) with existing or future technologies can pose a problem years after the generation or migration of a digital file. While files may be compatible with the existing environment at the point of generation, any small update to this environment may render the files inaccessible. At this point the choices are clear: fix the file (possibly many thousands) or fix the technology stack (which may be a slow process with certain commercial hardware/software).

© DAVID consortium: all rights reserved

Ideally, the choices made at the point of creation would render a file 'compatible' for its lifetime (i.e. for the duration of the format generation). The answer here would appear simple: standardisation. If content producers, content consumers and vendors agree on a common set of formats, files can be guaranteed to be compatible with tools that support these standards. Furthermore, tools from various vendors can be operated synergistically.

Currently, the digital archive market is settling on a small number of essence encodings (MPEG-2, JPEG2000, FFV1) and wrapper formats (MXF, Matroska, Quicktime MOV). However, even these standardised and popular formats, with support from the community of users and vendors, create problems. For example, regarding MXF, AVID reports that "successful MXF interchange between two products depends on the relative compatibility of their MXF implementations. But interoperability may also depend on other factors, including essence compatibility and metadata compatibility. So MXF is not a panacea" [Avid, 2006].

Profiles can help to restrict format ambiguity and promise to improve compatibility. For example, MXF specifies a number of Operational Profiles (OP). However, AVID reports that "files created by products from different manufacturers may vary significantly in their structure and contents, even if they comply with the same Operational Pattern specification." [Avid, 2006]. Again, we find that profiles are not an easy solution to the compatibility/interoperability problem, but they are a reasonable approach.

If particular profiles are to be used, validation and normalisation must also be used. Validation ensures that files conform to the specification, e.g. that an MXF file header and essence agree on the encoding. Normalisation ensures that content in disparate formats is converted to one of a set of agreed-upon profiles, thereby reducing the overhead of dealing with multiple profiles. However, normalisation on ingest introduces another conversion step and, therefore, could have a negative impact on quality.

Front Porch Digital, founder members of the SMPTE working group on the Archive eXchange Format (AXF) standard [ST2034-1], claim that a particular format "supports interoperability among disparate content storage systems and ensures the content's long-term availability no matter how storage or file system technology evolves" [OpenAXF, 2011]. Future proofing content in this manner involves including more contextual information in the file itself. OpenAXF includes a kind of file system within the file, so that it is self-contained, self-describing and can abstract the underlying storage technology.

Obsolescence

At some point in the future, it is very likely that every much-deliberated-over 'preservation' file format and accompanying technology stack will fall into obsolescence [Rosenthal, 2007]. Large parts of archived contents will be rendered inaccessible if support lapses for a particular flavour of digital format chosen by the archive.

Formats, tools and equipment can all become obsolete, requiring great effort in reverse engineering even if the specifications are open and publicly available. The strategies to reduce the risk of obsolescence are in careful selection of the tools and formats, and in diversification. To deal with obsolescence pragmatically, archivists tend to migrate content to new (and hopefully long-lived) formats. Alternatively, emulation of the underlying (obsolescent) technology can provide a constant environment in which to access content in otherwise obsolete formats [Gledson, 2010].

Selecting a format that can be guaranteed to have active tool support for decades is a very difficult task. Good strategies involve choosing a format with wide and active community support and public specifications and, ideally, open-source implementations [Rosenthal, 2009b]. The choice of format for preservation master copies can be specific to the archive, in which case it is recommended to perform an analysis as to the suitability and longevity of the format. Graf and Gordea have proposed a risk analysis for choosing file formats for preservation [Graf, 2013]. Alternatively, the US Library of Congress [LoC] and UK National Archives [Brown, 2008] have published sustainability criteria for formats, which can be used to inform the decision.

© DAVID consortium: all rights reserved

A diverse choice of technology should ensure that the risk of obsolescence is minimised. Selecting different storage devices/mechanisms, from different manufacturers, and using different formats reduces the risk of wholesale loss owing to dependency on a single technology that loses support in the future. Diversification addresses the problem of obsolescence in that the product life-cycles, adoption and support vary for different devices/formats from different vendors/communities. However, diversification increases cost and complexity, as the archive must maintain multiple storage stacks and preservation 'recipes', possibly with varying periodicity in their generations (as some formats become obsolete before others). This increases the management and maintenance overhead, so there is a trade-off to be struck.

The risk of obsolescence is reduced by a policy of diversification within a single archive, but there are benefits, at a macroscopic scale, to diversity within the preservation community. For example, the risk increases if the whole marketplace adopts a single solution (e.g. a single tape robot vendor or the same MXF-JPEG2000 implementation).

Redundancy

A tried-and-tested strategy for recovering from data loss is to keep additional copies of the data. Redundancy can be introduced at many levels of a preservation system, i.e. within the file format structure, at the application layer, at the operating system layer, and at the storage device layer. The safest option is to employ redundancy at several layers of the stack, essentially designing for failure.

Many systems requiring high degrees of data safety use redundant devices that manage data replication according to a scheme. RAID arrays offer a number of ways of managing data replication across devices. Alternatively, replication can be managed at the application layer using a set of low-reliability storage devices that can be quickly replaced. This strategy is often referred to as JBOD (Just a Bunch Of Disks [Rouse]).

In combination with redundancy of devices, diversification helps to improve robustness. Failure modes should be different for different devices, different vendors, even for different batches of the same model. This protects against systematic failure that could take out a large part of an archive, but does not protect against random failures.

Random failures can also be dealt with using redundant data. Erasure coding, CRCs, cryptographic hashes (e.g. MD5, SHA-1), on the whole or partial file, are all examples of data redundancy that can be used to detect data loss and aid data recovery. Redundant data can be used in the storage layer, the file system, the application/database layer, or within the file format. JPEG2000 uses a 'byte stuffing' technique which aids error resilience [Bilgin, 2003]. MXF duplicates essential metadata in several partitions. Diverse, redundant ways of representing data in a file format, such as offset-based timing and marker-based timing, also give multiple opportunities to recover if one index is corrupted.

File-level replication, whether it be handled by the application, or manually by the user, is a common form of redundancy. Bit safety should be theoretically dealt with at the carrier level, e.g. through erasure coding as mentioned above, not least because these reduces management and complexity at the application level, but also because erasure coding has been shown to have better MTTF than replication with similar storage/bandwidth requirements [Weatherspoon, 2002]. However, file-level redundancy is tempting as it is analogous to 'copies on shelves' and gives the user visibility of the digital copies (and possibly control over their location) and therefore confidence in the replication strategy.

Replicated content may need to stay within the purview of the archive (even if some copies are stored off-site), but some content can be distributed to other institutions, provided that rights can be effectively controlled. Stanford's LOCKSS ("lots of copies keeps stuff safe") approach [LOCKSS] uses decentralised storage to improve data safety and availability.

© DAVID consortium: all rights reserved

Simplification

Many of the above-mentioned strategies warn against increasing complexity in the pursuit of improved robustness. Simplicity should be the watch-word in preservation systems, not only for current operations, but also so that the technology can be understood in the future, when content must be accessed, but the expert knowledge is out of date.

Using simple file formats and simple profiles is a good step towards not exposing the archive to the risk of ambiguous or 'dark' metadata, which may cause problems in interpretation in the future. It is often noted that the archive must consider the 'burn line', i.e. the amount of technology (and knowledge) that can be lost while the data remains accessible (and meaningful). This should preclude deep integration with a particular preservation system, without which the data is unusable, as it relies on an arcane database schema. Approaches to avoid such problems are to use file formats that are 'self-describing' or 'sidecar' files containing metadata in a well-described format (e.g. XML).

The MXF standard allows data in the file to be partitioned; the size of the partitions is chosen arbitrarily by the encoder implementation, in order to flush the data and write an index segment. More partitions means greater file complexity, as the index segments are spread throughout the file.

In addition to file complexity, one must consider the complexity of the storage. For example, the BBC chose a simple approach when archiving to tape [Cunningham, 2007]. No spanning of tapes was allowed, so only complete MXF files were stored on any single tape. This reduces the risk of retrieving only a partial file.

Control over storage must also be considered, as remote storage (e.g. cloud storage) might increase the risk to which data is exposed, or might be a safer solution, depending on the processes in the archive. Remote storage is often considered as it reduces capital expenditure and allows storage to be scaled according to the needs of the organisation. This argument is valid when the organisation cannot afford much capital outlay (e.g. start-ups) or when storage requirements will change rapidly and unpredictably. The storage requirements in an archive are often predictable and some analyses suggest that cloud storage is not cost effective for digital preservation [Rosenthal, 2013].

7.1.2 Robustness of People

Technology is only part of the problem. There is a great likelihood that any loss event can originate from human intervention (or the lack of it). People are part of the preservation system and introduce errors, either unwittingly or maliciously, or through inaction fail to prevent the system reaching a state that leads to loss. The problem is then to ensure that people are aware of the indicators of loss, so that they can detect them, and know the process to follow to prevent or correct damage that may occur.

Understanding and modelling human error is a notoriously difficult task. Models of problem solving typically reduce to observation of patterns, a decision making process, in which a particular course of action is selected, followed by application of the solution in a feedback loop. Therefore, it is important to capture and communicate knowledge relating to the recognition of characteristics indicative of the onset of damage and identification of which solution to apply.

Standard operating procedures and guidelines can codify captured knowledge and provide a reference for operators. These can be specific to the archive, or can make use of ad hoc, community-based efforts, such as The A/V Artifact Atlas [AVArtifactAtlas]. Operators must be trained in the use of technology, so that it reduces the likelihood of misconfiguration or misuse of the tools. It helps to use standard, as opposed to bespoke, solutions where possible, so that operators have a wide and common source of expertise on which to draw.

© DAVID consortium: all rights reserved

7.1.3 Robustness of Processes

Preservation systems comprise operators and technology; the former use the latter as part of a set of steps, forming a process or workflow. Such processes are susceptible to errors of omission, commission, and mis-ordering of the steps, which can cause unexpected results, including damage to AV content, which may not be detected until much later. Furthermore, failure of an individual step in the process, if not detected and rectified, can propagate to later steps in the process, which again can result in damage to content. Strategies to improve the robustness of preservation processes involve careful specification of the process, quality checking and management of changes to the process.

Quality checking

Besser notes that digital preservation has thrown the focus on *active* management: "the default for digital information is not to survive unless someone takes conscious action to make them persist" [Besser, 2000]. Active management involves checking (automated or manual, or a hybrid of both) both of the process (i.e. that the steps are being followed correctly) and that the steps are producing acceptable output.

Output must be checked at the bit level, the format level (i.e. the wrapper and essence), and at the content level. Checking can be put in place at any point in the process when assets are created, processed, or accessed, as well as periodically during the asset's lifetime.

Integrity verification (also known as scrubbing or fixity checking) is the most basic form of QC, which verifies that a file has not changed. Typical methods are to compute a checksum (such as MD5 or SHA-1), which can be stored and verified against in the future. It is possible to generate sub-file checksums, e.g. on each frame, or on the frame video and audio content, which helps to identify the area of the content that is damaged.

Format-level QC aims to verify whether a file fits a particular profile or standard. This is important so that tools can process the file in the future. Systems such as Interra BATON [Sumanta], include file-based QC methods that verify that files are free of error and will play out correctly.

Quality checking the content of AV assets is important when ingesting and migrating content. Migration from one format to another can introduce damage if not monitored appropriately, such as this example of over one million digitised newspaper pages transferred from TIFF to JPEG2000, some of which were truncated owing to a faulty migration process [Wheatley, 2012]. QC can be achieved through labour-intensive manual inspection or through (semi-)automated QC tools, such as VidiCert [VidiCert].

The XCL project has created tools to ascertain whether a migration process has correctly preserved all the 'significant characteristics' in the target format [XCL]. The approach involves describing source and target file formats in a common format (XCL), so that instances of files in the original format can be directly compared to the converted file. This approach is currently only defined for document formats (e.g. MS Word and PDF) and static images.

QC is not a perfectly reliable method for detecting damage. If damaged content slips through the QC process, it can be iteratively improved (with the benefit of hindsight) so that the particular variant of damage will be detected next time; however, it is very difficult to determine in advance what to check for. The level of QC fidelity depends on the intended use of the content and, as always, there is an inevitable trade-off with cost/time.

Change management

Any change to the process can have sometimes unexpected and dramatic effects on the output. Upgrading the tools, swapping the people involved in parts of the process, changing the steps or order of steps in the process, or a wholesale change of the process must be checked and their effects verified.

© DAVID consortium: all rights reserved

Subtle changes to one tool in a chain can have far-reaching implications, which may not be detected until it is too late. It is typical to have a standard set of test material that can be used to 'regression test' changes to a workflow; however, this may not always be possible if the workflow is entirely new, such as a digital migration workflow that replaces an analogue workflow.

Redundancy

Redundancy can also be useful in improving process robustness. Using parallel processes (possibly using diverse methods) can provide several outputs that can be cross-checked to catch errors in one of the processes. Naturally, this increases the amount of resources required to process the same volume of AV content.

Simplification

Again, simplicity is paramount in the design of effective processes. Processes that are complicated are difficult to understand, difficult to follow and difficult to analyse to determine optimality and likely points of failure.

Provenance / audit

Given that an asset will undergo many movements, changes and transformations in its lifetime, it is essential to understand the chain of events that has led to damage, if it should be detected. Provenance management and auditing of a file's history, in terms of the operations performed on it, are part of good preservation.

The chain of operations should be recorded in a way that can be audited when required. The record should also include failed operations, which required reprocessing or rollback of state. Often system and application logs can shed some light on the history of a file, but these are rarely available or complete in the long term. Some preservation systems keep records of events in a structured format, such as PREMIS [PREMIS]. In some situations, the history of a file must be reverse engineered on a case-by-case basis from inspection, e.g. by using tools developed in the REWIND project [REWIND], which aims to detect tampering through re-encoding and can in some cases detect the original codec or original camera used to produce the content.

Often the fail-safe strategy is to keep the original source of the AV content. As migration processes improve over time, migrating from the original source could give better results than migrating from an intermediate format [Lacinak, 2010].

7.2 IT-based preservation at INA

Within INA, IT-based prevention of digital AV loss relies on a number of different strategies, tools and workflows, complementing each other. Those strategies rely mainly on the security of AV media files on physical storage, on the existence of three different copies of each media file, and on the security of the systems as a whole. This is detailed below, focusing on the Inamediapro commercial contents delivery workflow.

7.2.1 Physical security of AV data

Security of physical AV media data storage is ensured differently, depending on whether storage is on data tape (LTO5), or on disk.

Security of data on the frequently-accessed 3000 tapes stored into the automated tape library (typically up to 300 loads per tape) is ensured by monitoring closely the error rate using 3rd-party software, automatically raising an alarm when the error rate rises above a specific threshold, and triggering a duplication and replacement of the tape in error when this happens. An additional security level is given

© DAVID consortium: all rights reserved

by forbidding writing on a tape when it is nearly full, which usually happens a few hours after it starts being written.

Security of backup tapes is ensured by storing into a separate remote location, under climate-controlled environment.

AV files that are stored on disks use the RAID 6 (double parity) strategy to prevent losses even in case of two disks failures in the same enclosure.

7.2.2 Backups of AV data

When an AV file is first generated, it first exists as a single copy in INA. However, this is very temporary; the workflow plans for three different copies, and is only complete when all three copies have been generated. The three copies are labelled as Main, Recovery, and Backup. Main copy for each file is currently stored on one or two data tape in the main tape library managed by a HSM system. Recovery copies are stored on a number of RAID 6 appliances in a remote location aimed at restarting the activity in case of a disaster. Backup is stored on a data tape and stored on shelves at another different location. Care is taken that files stored on the same Main data tape are also stored on the same Backup tape, in the view of limiting the number of Backup tapes to be accessed, should several Main tapes be lost. Locations (tape ID, RAID ID and path) of the three different copies of each file is known and maintained in a database.

In some cases a fourth copy of the file exists, e.g. when the file was present on the earlier data tape version (LTO-3). Although not strictly speaking a file, INA also keep on shelf for a large part of the AV contents a higher quality Archive Master copy (usually a DigiBeta tape) that can be used to re-create a file if needed.

7.2.3 System security of AV data

At the system level, all the usual IT strategies for ensuring the continuity of service are used: databases are backed-up every day, systems are duplicated, revision control is used... Databases are used to track the location of each AV file, and verification tools are used to keep those databases up to date, and correct locations errors if any. In the event of a large subsystem failure, a Disaster Recovery Plan was set up, with the objective of restoring the core capacity for delivering contents to customers in less than 48 hours. This plan relies on the replication (several Petabytes), in a remote location, of all the files (“Recovery” copies) and other data, software, and servers, needed for re-starting the day-to-day activity. This Plan has never been fully triggered yet, but “Recovery” copies are accessed when needed for restoring “Main” copies.

7.2.4 AV data Security overall in INA

As a result of the combined strategies exposed above, none of the several million AV files used in commercial activities was ever lost in INA. However, it does happen from time to time that a file is found to be flawed from the origin, e.g. when DigiBeta head-clog during ingest has gone un-noticed. In that case, the solution is usually to re-generate the file from the Digibeta copy, when exists. In the worst case, content has to be re-digitised.

Encoding and wrappers compatibility problems are however quite frequent, and although they usually do not result in “losses”, they can introduce bottlenecks and delays. Examples include wrong or inconsistent pixel or picture ratios, timecode problems, soundtrack problems... It is tried to avoid such problems by preparing as complete specifications as possible, and detecting the problems at the Quality Control step when the files are generated, but, quite often, the problems go unnoticed until a new exploitation scenario is implemented (e.g. a new editing tool, or a new export mechanism is used). Mitigation procedures depend on the age and number of files affected. Recent files are re-generated or patched. When numerous older files are at stake, it is often easier to adjust the tools and procedures than to modify the files, to allow using the affected files anyway.

7.2.5 Comparison with analogue workflows

The situation in the digital worflows is very different from that of the analogue contents, where the number of copies for each programme, their condition and playability vary greatly (and usually decrease with time due to media ageing).

The systematic multiple copy strategy used in the digital domain was only applied to analogue contents in INA for very specific case, when it was clear that a specific part of the collections was endangered.

© DAVID consortium: all rights reserved

This was the case for example, for nitrate film material, where a safety film copy was made on the long run for most of the nitrate collections. It was also the case for SepMag (separate magnetic track), the soundtrack of a large part of INA's 16mm film collections, where a massive digitisation plan was started when it was confirmed that the Vinegar Syndrome was progressing at a fast pace.

INA maintains a database of physical media, this database maintains the filiations between media, it is therefore theoretically possible to devise the best medium to start digitisation from, but costs considerations also have to be taken into account. Therefore when preparing digitisation plans a large number of criteria are considered. Those result in batch lists that are sent for processing and digitisation. When digitisation is made, this results into an Archive Master copy (Digibeta tapes for video, HDCAM-SR for film, progressively being replaced by MXF-JPEG2000 files). The Archive Master copy is the entry point into the IT-based preservation workflows in INA.

7.3 IT-based preservation at ORF

At ORF we have started in mid-2012 with pure file-based archiving; since then all production-units at ORF has been (or will be until mid-2014) switched over to a "file-only"-digital workflow. All production-units, which already changed to this "tapeless"-workflow, preserve their content in our new file-based storage system (called ESYS, which is short for Essence-System). ESYS is triggered and managed by our main TV-CMS "FESAD" (a collaboration-based system from the ARD). Nucleus of ESYS is an IBM-TSM Storage (LTO5-based) managed by a customised AREMA (IBM) System (formerly ADMIRA).

Already in the early planning stages, the topics of how to avoid loss and how to recover from damage and "wrong decisions" took an important role in our discussions. At that time the continuation of the old rule from analogue times "Always keep the original source" was taken up for all migration issues; for the "born digital" or "file-born" content this was no longer valid, since the file-storage of modern acquisition units (cameras, etc.) is not meant to be kept any longer as an archive-carrier. So in combination with the needs of a 24/7-availability of the new ESYS-system the concept of "100% redundancy + offline-copies" was born and also implemented: all parts and sub-systems of ESYS, including the LTO-storage, are built up completely redundant on two different sites; all content is stored equally on both sites (those sites are also in different seismic zones, etc.). Therefore even a total shutdown at one site would not affect the availability of the system (only the available bandwidth). A third copy on LTO is stored offline at a third site (ORF-Centre) for additional security and legal issues.

During the final planning stages of our Digital Migration Project (DiMi) we had to decide to no longer stick to the old rule of "keeping the source", since due to commercial reasons the space will be needed otherwise. So after the final checks in the DiMi (see also 8.3) all analogue (and digital) sources will be disposed of.

In addition to the Quality Check-Routines in ESYS (automated; general check of incoming files; we check conformance of files against profiles to prevent inconsistent file structures in our archive; actions: on negative results the file is rejected and the source-system is alerted) and TSM-internal routines of file-recovery (check of read-errors; internal error-statistics), there are widespread additional QC-routines already installed or in their final implementation stage. QC starts at ingestion-level (new file-based material is tested thoroughly and divided into three different categories:

Class1: The source of the files is known, no detailed testing is performed.

Class2: Files with known general file format but from non-trusted sources (mainly XDCAM-HD material, but different flavours) Intention: The file structure must be computable by the broadcast infrastructure, else a re-wrapping or transcoding process make the files useable.

Class3: Unknown file sources: an automatic, or if not possible, also manually executed transcoding tries to make the files useable.

This approach minimizes the need of full rejection of files and ensures the availability of material/content in the production-process), continues during production (due to legal and content-based quality-issues main parts are still manual here and are mainly focusing on content-related quality issues) and finally before broadcasting (again integrity and compliance checks; on fail the file is rejected and the

© DAVID consortium: all rights reserved

production-unit is alerted) and archiving (see ESYS QC-routines above). Unfortunately some powerful tools on the market (like Baton, etc.) are still not very useful for large amounts of material and daily workflows, although currently used for that purpose here. ORF expects major improvements in this sector by the tools to be developed in DAVID.

So while the “Redundancy-Concept” of ESYS is quite similar to the old “Keep all copies”-Rule (and sometimes even better, since now for ALL content is redundant, which was not the fact so far), we do expect a similar or higher rate of “Preservation-security” for our contents in our new ESYS. The automated routines in TSM etc. should also work better than the old “check on use” done with the tape-based video, providing constant control and avoiding “mass effect” errors in the future.

For the QC-routines (implemented and planned) it is important to point out that they are more numerous than those in the “old tape-times” and (at least this is the expectation) more reliable, as they are mostly automated, while the earlier ones where always carried out manually.

In summary the overall expectation here at ORF is that our new file-based environments for production and archiving in the TV-domain (ESYS, etc.) have lower risk-levels for total loss and/or damage, which are comparable to the risks in the analogue tape domain (due to higher redundancy-levels); unfortunately the file-world brought along new risks and threats, based on format-, codec-, and wrapper-incompatibilities and the permanently changing soft- and hardware environments/systems. To address those new risks the workload in tasks like “Securing QoS (Archive-Services in total, including preservation and access) and QoF (Quality of File)” is more and more transferred to the development/implementation/evaluation-phases and away from the daily workflows. This fact bears but another risk: that some problems/errors are kept undiscovered for a long time and that the amount of affected content is therefore very large. As a reaction, the QC-routines have to be permanently adapted and updated to provide secure workflows and systems with no loss and small damage.

Additional Information on DiMi: to gain as much as possible of the positive effects of a file-based production- and archive-environment (lower risk on loss and damage from a higher redundancy-level; faster and easier access on archive-content; etc.), ORF will migrate approx. 360,000 hours of AV-content from IMX and Digital Betacam to MXF D10 OP1a; those will be stored then in the ESYS-system (see also section 8.3).

© DAVID consortium: all rights reserved

8 Conceptual Risk Management Framework

8.1 Archive risk management

Risks, as defined by ISO 31000 [ISO31000, 2009], are the effect of uncertainty on objectives. In the context of DAVID, uncertainty arises from random or systematic failure of preservation systems and processes, the effect of which is to cause damage to AV content.

Risk management involves identifying, assessing and prioritising risks, such that appropriate risk treatment strategies (as described in Section 7) can be applied to avoid, mitigate or recover from the effects of the risk.

8.1.1 Risk management in safety-critical engineering

Many risk modelling and management techniques have been developed in the safety-critical sector (e.g. the design and construction of power plants, aircraft), in which failure of a component part can cause injury or loss of life. Risk management is also prevalent in financial systems, in which failure can have great financial impact.

Safety-critical risk management focuses on detection and control of a 'hazard' event, i.e. that state of the system in which normal operation will lead to injury or death with some probability. Understanding how hazards arise, and what their effects might be, has been well studied, and many techniques have been proposed for their analysis. The approaches can be distinguished as inductive ('forward-looking') or deductive ('backward-looking').

Among the set of deductive techniques are Ishikawa diagrams, which aim to explain the contributing factors to a loss event (e.g. an accident) using distinctive 'fish-bone' diagrams. The causal factors analysed are typically categorised and include: People, Methods, Machines, Materials, Measurements and the Environment.

Other deductive techniques recognise that combinations of faults must occur together for the risk to eventuate, while other risks can occur as the result of distinct and independent faults. Fault Tree Analysis is one such approach to combining faults using logical operators (AND, OR) to create a tree, the root of which is the loss event under investigation.

James Reason's 'Swiss cheese' model [Reason, 2000] aims to describe how circumstances align to 'allow' a failure to occur. This approach is motivated to describe failures arising from human factors. It describes systems and individuals as layers of 'cheese' in which the holes are individual weaknesses. An accident occurs when all of the holes momentarily align, allowing a hazard to pass through the layers of defence that normally catch it before it becomes a problem.

Inductive techniques start from a system description and aim to determine what could go wrong. Failure Modes and Effect Analysis (FMEA) analyses the system and its components to determine the way in which they can fail and, through expert judgement, to ascertain the likely effects of such failure. In this way, an analyst can determine how (if uncontrolled) failure can propagate throughout the system. Control mechanisms can then be put in place to mitigate the failure modes.

In Hazard and Operability Studies (HAZOP) an analyst takes a process flow description and, using a list of guidewords, perturbs the flow to determine the likely effect of flow failure. Guidewords commonly used in HAZOP include early, late, omission, commission, reverse, etc. Each guideword is applied to the parts of the flow and, through expert judgement, the effects of missing a step, performing a sequence of steps out of order, or later than required, is determined.

© DAVID consortium: all rights reserved

8.1.2 Financial risk management

Financial risk management aims to minimise a firm's exposure to risk using financial instruments. Controlling risk in this way aims to reduce the likelihood of loss of economic value. Financial risks typically stem from market risk (uncertainty in the future value of stocks and shares), and credit risk (uncertainty in creditors' ability to pay their debts). Risk measures are used to describe, for example, the probability of a creditor's defaulting and the expected loss that this would generate, as well as the probability that a given value of a portfolio is lost owing to market changes.

Risk measurement techniques, such as those described above, rely on knowing the value of investments and loans, which, being monetary in nature, is relatively easy to assess. However, being able to estimate the value of such assets as digital AV content, is a much more difficult task. Using such a value-based argument to motivate investment in preservation systems seems obvious but is not as easy as it sounds. How do we appraise the value of an AV asset and the degradation of value that damage causes? In his book "Appraising Moving Images: Assessing the Archival and Monetary Value of Film and Video Records" [Kula, 2002], the author approaches the philosophy of 'monetary appraisal' when considering whether to preserve AV assets. While the focus in the book is on selection of analogue content for digitisation and preservation, the guidelines are equally applicable to on-going selection for content migration. This approach asks archivists to base their decisions on a work's 'archival value, i.e. its value as an historical record, and emphasises that the work must be understood in context.

Archives have, among others, started to invest in preserving digital content on the basis that an archive of digital AV material can 'pay for itself', as digital content is more easily monetised than analogue content through improved access [Comité des Sages, 2011]. The value is therefore related to the price the market is willing to pay to get access the content. Kaufman investigates business models for revenue generation by exploiting AV assets [Kaufman, 2013]. While some of these business models have been successful and others will no doubt be developed, constructing a 'return on investment' (ROI) argument for preservation is a difficult task.

In contrast to ROI, Kara van Malssen recommends using a 'cost of inaction' (COI) argument [van Malssen, 2013] to stimulate investment in preservation. Chris Lacinak has developed a tool that shows how much content would be lost (assuming an exponential decay of media) if content is not transferred in time [AVPreserve]. If coupled with monetary appraisal of the content, such tools would allow archivists to argue the value put at risk of loss. Such loss-oriented arguments can be very effective in stimulating additional investment in risk treatment strategies. Risk management using measures such as expected loss and value at risk will be investigated further in section 8.2.

8.1.3 Archive risk management

Current archives, charged with preserving AV content for future access, typically deploy a number of the strategies for avoiding, preventing or recovering from loss that have been introduced in section 7.1. Specific examples of this were given for the French national archive, INA, in section 7.2, and for the Austrian broadcaster, ORF, in section 7.3.

These archives are engaged in a process of long-term digital asset management (DAM) [Green, 2003], specifically media asset management (MAM), which focuses on storing, cataloguing and retrieving digital AV content. Tools exist to support the MAM process, such as the open-source tool DSpace [DSpace], some of which support the risk treatment strategies identified above. However, these tools do not include a model of risk. The archive must decide on risk indicators and define the way in which these can be measured in order to monitor them, often using separate tools to do so.

Some MAM tools conform to the Open Archival Information System (OAIS) reference model [OAIS], which defines the concepts and framework for long-term preservation. The OAIS model recommends that periodic risk analysis reports be created as part of preservation planning.

© DAVID consortium: all rights reserved

If an archive provides a service to content producers, ISO 16363 sets out specific audit guidelines, such that users of the archive can be assured that it is a Trusted Digital Repository (TDR) [ISO16363, 2012]. The standard defines the attributes (including compliance with OAIS) and responsibilities of a TDR, such that it can be certified as such. Compliance with the standard, by fulfilling the attributes and responsibilities, is aimed at reducing the risk to which the repository puts the content it holds, but the standard mandates no particular risk management method.

Preservation planning tools help archives to understand and investigate the risk involved in different proposed preservation solutions. Many of these tools are only just emerging from early research prototypes, such as Plato [Plato], iModel [Addis, 2010] and the aforementioned beta COI tool from AV Preserve. These tools look at the growth in content, cost of storage and patterns in storage reliability in order to determine the risk of loss. However, calibration of these tools is essential if the results are to be useful. While reliability statistics are difficult to determine for some existing storage systems, predicting future growth of data storage capacity, requirements and reliability is almost impossible. Some analyses [Addis, 2013] predict that as ever more 8K AV content is ingested into archives, the growth in data volumes will, with all likelihood, outstrip the growth in storage capacity and increase in data write rate, such that it becomes impossible to store and replicate all content as it is produced.

While whole preservation planning tools aim to balance the cost of preservation versus the risk of content loss (or damage), specific risk assessments can be carried out into particular aspects of the AV assets. The Simple Property-Oriented Threat (SPOT) model has been proposed as a model against which to evaluate the effectiveness of a preservation strategy to maintain an asset's essential properties. The essential properties have been derived from an extensive review of literature and are: availability, identity, persistence, renderability, understandability, and authenticity of digital objects [Vermaaten, 2012]. The SPOT essential properties are described in greater detail in the proposed risk model in section 8.2.3. At the iPRES 2012 conference, it was proposed to use an analysis of PREMIS metadata to close the loop between preservation planning and operations to show whether risk treatment strategies are effective [Lavoie, 2012]. Data is captured and analysed using the SPOT risk model as part of the Preservation Health Check pilot [van der Werf, 2012].

The above tools do not take into account the specific file format of the digital content, the longevity and interoperability of which presents a significant preservation risk. Work on the risk analysis of file formats ranges from early investigations in 2000, using pair-wise conversion between document formats in order to derive risk measures from the number of errors in a set of test files [Lawrence, 2000] to recent work in 2013, in which risk factors are automatically derived from linked open data for different static image formats [Graf, 2013]. It is clear that an effort is required to incorporate file format risks into the cost-versus-risk-of-loss-based planning tools.

8.1.4 Workflow/business process risk modelling

Workflows are often used to describe business processes and, increasingly often, are used to automate some or all of the process. Automated workflow execution is possible if the process is specified in a machine-interpretable fashion, such as using BPMN [BPMN]. As described in section 8.1.1 with respect to HAZOP, risks are inherent in processes, as individual steps may fail, causing consequences for later parts of the process, or if the process is not executed correctly. Risk-aware business process management is critical for systems requiring high integrity, such as archives.

A good and recent review of business process modelling and risk management research is given in [Suriadi, 2012]. Risk-aware business process management has several parts:

- Static / design-time risk management: analyse risks and incorporate risk mitigation strategies into a business process model during design time (prior to execution).
- Run-time risk management: monitor the emergence of risks and apply risk mitigation actions during execution of the business process.
- Off-line risk management: identify risks from logs and other post-execution artefacts, such that the business process design can be improved.

© DAVID consortium: all rights reserved

Several approaches have been proposed to model business processes and risk information such that it enables risk analysis. Rosemann and zur Muehlen propose integrating process-related risks into business process management by extending Event-driven Process Chains (EPC) [Rosemann, 2005]. Risks are classified according to a taxonomy including structural, technological and organisational risks.

Analysis of process risks is difficult given that operational risks are highly dependent on the specific (and changing) business context. Many risks are caused by business decisions (e.g. preservation selection strategy, migration path), so large volumes of data required for statistical methods are often not available for analysis. Those who subscribe to this thesis use structural approaches, such as Bayesian networks, influence diagrams, and other techniques introduced in section 8.1.1. For example, Sienou et al present a conceptual model of risk in an attempt to unify risk management and business process management using a visual modelling language [Sienou, 2007].

In contrast to the above thesis, some believe that run-time analysis of risks is possible with a suitably instrumented execution process. Conforti et al propose a distributed sensor-based approach to monitor risk indicators at run time [Conforti, 2011]. Sensors are introduced into the business process at design time; historical as well as current process execution data is taken into account when defining the conditions that indicate that a risk is likely to occur. These data can be used for run-time risk management or off-line analysis.

Given that analysis of business processes using structured and/or statistical approaches can reveal vulnerabilities, it is important to control the risk that these vulnerabilities lead to loss. Bai et al use Petri nets (a transition graph used to represent distributed systems) and BPMN to model business processes and to optimise the deployment of controls, such that the economic consequences of errors (measured as Conditional Value at Risk - CVaR) are minimised [Bai, 2013].

The PrestoPRIME project described in BPMN the preservation workflows that were implemented in the preservation planning tool iModel [iModel]. It is clear that tools are required to model (in a flexible way) the preservation workflows and annotate them with risk information.

8.2 Risk Management Model

8.2.1 Risk management steps

In DAVID project we propose a framework with the following steps for archive risk management model:

- Identify workflows and processes taken place during digital preservation
- Define objectives of risk management for digital preservation in archives
- Identify risks/threats in workflows/processes and any negative consequences for individual tasks and overall workflow
- Identify available procedures dealing with risks/threats and any associated known costs and time
- Analyse risks/threats
- Classify risks/threats according to SPOT model
- Define scenarios which are of interest to the user
- Select the most suitable risk management model
- Simulate scenarios and evaluate risks according to selected model(s)

This framework will allow to identify the objectives of individual archives, analyse the risks/threats which affected digital preservation process in a given archive, discover generic and specific features of digital preservation, and suggest risk measures which suit the given archive and its goals.

© DAVID consortium: all rights reserved

8.2.2 Risk measures

Three of the most popular functions measuring risk are proposed for preservation process at this stage in DAVID: Expected loss [Berger, 1980], Value at Risk (VaR) [Duffie, 1997] and Conditional Value at Risk (CVaR) [Rockafellar, 2000]. All these functions have advantages and drawbacks, easy to calculate if no control is used, and allow incorporating control procedure with different level of difficulties. The idea to use Expected loss, VaR and CVaR for risk management in digital preservation was inspired by Bai et al [Bai, 2012]. They used these risk measures to assess economic consequences for an order-fulfilment process of online pharmacy. Moreover, in [Bai, 2012] the authors took into account how the topological structure of the process can have its effects on error propagation and risk mitigation. Here we will follow their procedure with some adaptations for digital preservation in archives.

Expected loss is the average magnitude of negative consequences which can occur in the digital preservation workflow. Expected loss is a satisfactory measure of risk when the loss can be viewed as normally distributed with a fixed standard deviation. The estimation of the Expected loss with no risk mitigation controls in place requires the following attributes related to the digital preservation:

- workflow of a given digital preservation process or sub-process
- probabilities or frequencies of following different paths after a decision points
- for each task
  - list of all possible threats that can take place
  - frequency or probability of each threat based on historical data or estimate
  - negative consequences of each threat in currency, level of severity or percentage

The estimation of the Expected loss without control for a given workflow is given by

$$\mathbf{E}(l) = \sum_{i,m} p_{im} s_{im} \sum_{j} \gamma_{ij}$$

where $\mathbf{E}(l)$ is the estimation of the Expected Loss

p_{im} is the probability that task *i* introduces the threat type *m*

s_{im} is the magnitude of expected negative consequence of type *m* threat

γ_{ij} is the threat propagation potential from task *i* to task *j*.

The threat propagation potential matrix is calculated as a sum of all k-step transition probability matrices

$$\mathbf{\Gamma} = \sum_{k=1}^{K} \mathbf{V}^k ,$$

where *K* is the longest possible path in the workflow,

V is a transition probability matrix

$\mathbf{V}^k$ is the k-step transition probability matrix.

If a given workflow includes controls and under assumption that all controls are used, the Expected Loss with controls can be estimated as

$$\mathbf{E}(l(x)) = \sum_{i,m} \left(1 - g_{im} \gamma_{+i} x_{im}^{a_{im}} \sum_{j} p_{im} \gamma_{im} s_{im} \right),$$

where $\mathbf{E}(l(x))$ is the Expected Loss with control,

x_{im} is the input level of resources at task *i* for threat type *m*; $0 \le x_{im} \le 1$ with $x_{im} = 0$ when no control is applied at task *i* for threat *m* , and $x_{im} = 1$ when all available control is applied.

a_{im} is the output elasticity of control resources and determined by available technology involved in control procedure. a_{im} defines the convexity or concavity of the effectiveness of control procedure.

© DAVID consortium: all rights reserved

g_{im} is normalisation factor that rescales the maximum effectiveness of a control procedure into the interval between 0 and 1.

If the Expected Loss is calculated with control, then the input level of resources at tasks for a given type of threat should be provided by the user together with some information about technology involved in control procedure.

As mentioned earlier, the Expected Loss shows only average loss during workflow operation under normality assumption. However, when loss distribution is skewed, then other risk measures such as VaR and CVaR are more adequate for the risk estimation.

VaR answer the question:

*What is **the minimum negative consequence** incurred in α% of worst cases*?

While CVaR answer the question:

*What is **the expected negative consequence** incurred in α% of worst cases*?

CVaR is always bigger than VaR, and depending on the purposes of risk management, the most appropriate risk measure will be selected. Figure 2 [Sarykalin, 2008] shows graphical representation of VaR and CVaR for better understanding these two risk measures. Reasons affecting the choice between VaR and CVaR are based on differences in mathematical properties, stability of statistical estimation, simplicity of optimisation, objective of risk management etc. Thus CVaR has superior mathematical properties and risk management can be done quite efficiently if optimisation is required,while VaR can provide better estimates for out-of samples if tails are not modelled correctly.

If the loss distribution is normal, **VaR** at confidence level *100×(1-α)*% can be estimated as

$$VaR_{1-\alpha} = -x_{\alpha} = -(\mu - z_{\alpha} \times \sigma),$$

where x_{α} is the left-tail α percentile of normal distribution $N(\mu,\sigma^2)$

μ and σ is mean and standard deviation of negative consequences s_{im}

z_{α} is the left tail α percentile of a standard normal distribution.

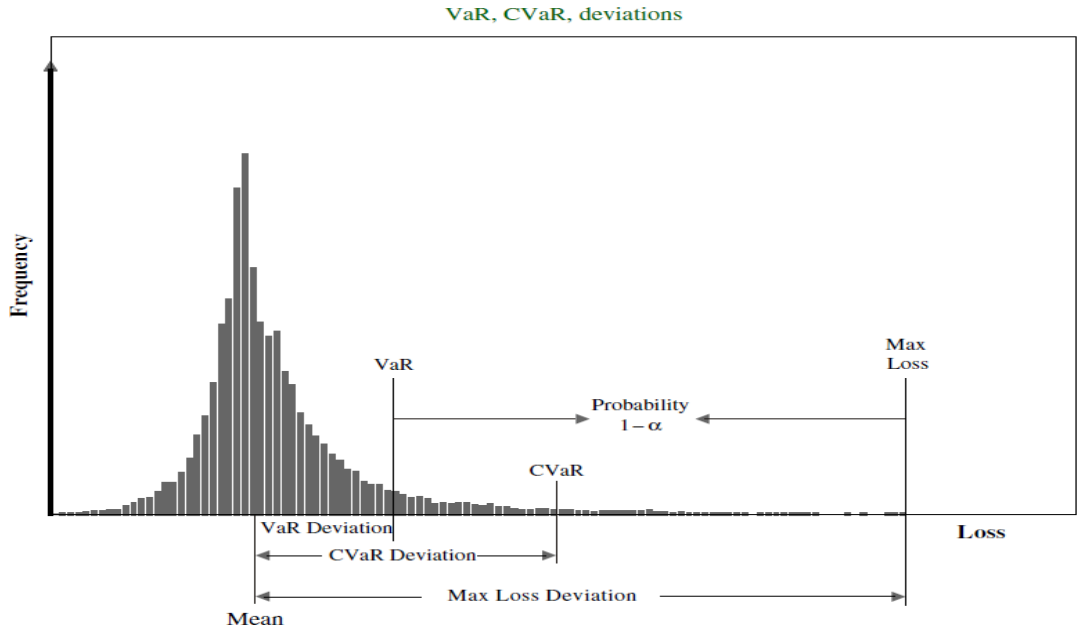

Figure 2: Risk functions: graphical representation of VaR, VaR deviation, CVaR, CVaR deviation, Maximum Loss and Maximum Loss Deviation

If the loss distribution is nor Normal, the VaR at confidence level *100×(1-α)*% can be estimated as

© DAVID consortium: all rights reserved

$$VaR_{1-\alpha} = -x_\alpha = -s_{im}(w),$$

where $s_{im}(w)$ is the w^{th} element of negative consequences sequence sorted in increasing order,

$w = \lfloor n\alpha \rfloor$ is the integer part of $n\alpha$, n is a number of negative consequences.

The main drawbacks of VaR measure are

- VaR does not show how serious negative consequences are beyond α percentile
- VaR is only computational tractable for normal distributions. If negative consequences are not normally distributed or have a finite number of scenarios, VaR is
    - non-smooth, non-convex and multi-extrema function

CVaR for normal loss distribution can be estimated as

$$CVaR_\alpha = -x_\alpha = -(\mu + k_1(\alpha)\sigma),$$

where $k_1(\alpha) = \left(\sqrt{2\pi} \exp\left(\left(erf^{-1}(2\alpha - 1) \right)^2 \right) (1-\alpha) \right)^{-1}$.

If the loss distribution is not normal, then CVaR is estimated as

$$CVaR_\alpha = -(\text{Average of all } s_{\text{im}} \le x_\alpha).$$

If control is applied, then both VaR and CVaR can be estimated based on negative consequences after control procedure.

The Expected Loss, Value at Risk and Conditional Value at Risk are not the only risk measures which can be used in digital preservation. The choice of risk measure will depend on the objectives of risk management for digital preservation and available information provided by a concrete archive. As DAVID project will progress, we will re-evaluate available risk measures and approaches and adapt our system depending on the user requirements, available information and goals of digital preservation at a given archive.

8.2.3 Classification of threats/risks in digital preservation

To classify possible threats in digital preservation, it is suggested to use the Simple Property-Oriented Threat Model (SPOT) for Risk Assessment. SPOT model [Vermaaten, 2012] defines six essential properties of successful digital preservation: Availability, Identity, Persistence, Renderability, Understandability, and Authenticity. We will not go in the details of describing the SPOT model here, however we will give a short definitions of each property and list threats associated with this property.

- ***Availability*** is the property that a digital object is available for long-term use. *Threats*:
    - A digital object deteriorated beyond restoration power
    - Only part of the digital object is available for preservation
    - A digital objects is not available for preservation due to disappearing, cannot be located or withheld
- ***Identity*** is the property of being referenceable. A limited amount of metadata is required for this property. *Threats*:
    - Sufficient metadata is not captured or maintained
    - Linkages between the object and its metadata are not captured or maintained
    - Metadata is not available to users
- ***Persistence*** is the property that the bit sequences continue to exist in usable/processable state and are retrievable/processable from the stored media. *Threats*:
    - Improper/negligent handling or storage
    - Useful life of storage medium is exceeded

© DAVID consortium: all rights reserved

- Equipment necessary to read medium is unavailable
- Malicious or/and Inadvertent damage to medium and/or bit sequence

- ***Renderability*** is the property that a digital object is able to be used in a way that retains the object's significant characteristics (content, context, appearance, and behaviour). *Threats*:
  - An appropriate combination of hardware and software is not available, cannot be operated or maintained.
  - The appropriate rendering environment is unknown
  - Verification that a rendering of an object retains significant characteristics of the original cannot be done (e.g. a repository is unable to perform sufficient quality assurance on migration due to volume)
  - Object characteristics important to stakeholders are incorrectly identified and therefore not preserved
- ***Understandability*** requires associating enough supplementary information with archived digital content such that the content can be appropriately interpreted and understood by its intended users. *Threats*:
  - The interest of one or more groups of intended users are not considered
  - Sufficient supplementary information for all groups of intended users is not obtained or archived
  - The entire representation network is not obtained or archived
  - Representation network of supplementary information is damaged or otherwise un-renderable in whole or in part
- ***Authenticity*** is the property that that a digital object, either as a bitstream or in its rendered form, is what it purports to be. *Threats*:
  - Metadata and/or documentation are not captured
  - Metadata maliciously or erroneously describes the object as something it is not
  - A digital object is altered during the period of archival retention (legitimately, maliciously or erroneously), and this change goes unrecorded.

Since not all possible threats/risks in digital preservation workflow will fall in the six properties mentioned above, we introduce extra possible state in SPOT model ***Other*** for such cases.

In the next section we will show how the proposed Risk Management Model can be applied to digital preservation on ORF workflow.

8.3 Analysis of ORF risk management

8.3.1 Identifying workflows and processes taken place during digital preservation at ORF

In this section we show how our framework can be applied to digital preservation workflow on example of DiMi workflow provided by ORF. At this stage of the project not all steps in framework were carried out, since the analysis of ORF risk management is ongoing process. DiMi workflow is given in Figure 3 and shows the full cycle of context preservation and digitalisation from selecting material from vaults to returning processed material back to the vaults and storing digital files.

© DAVID consortium: all rights reserved

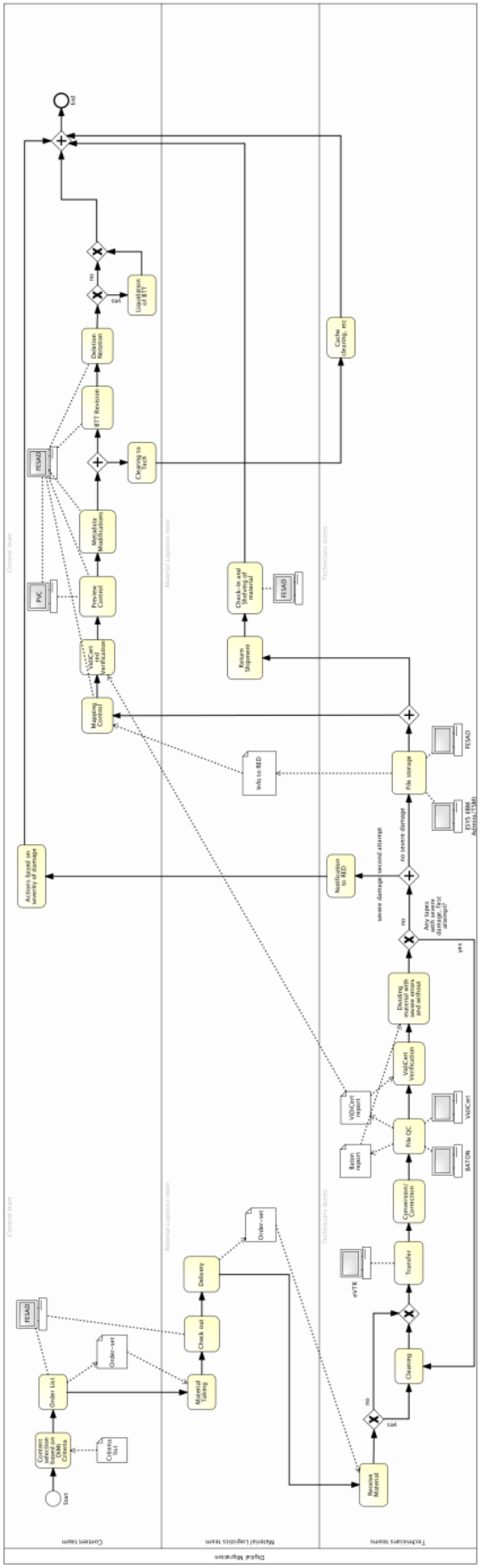

Figure 3: DiMi Workflow model

© DAVID consortium: all rights reserved

8.3.2 Defining objectives of preservation process and risk management

To understand the goals of preservation process and risk management for DiMi workflow the document-questionnaire was sent out to ORF. This questionnaire covered the following topics: objectives for risk management, consequences of something going wrong for overall preservation process, identification of risks/ threats in DiMi workflow for each task and their negative consequences on the following task, remaining workflow and overall workflow, cost and time associated dealing with risks/threats. Further we will discuss all these issues related to DiMi workflow. Please note the questionnaire is a dynamic document and more information will be added or updated during the project. These changes have to be taken into consideration when selecting risk measures and strategies for dealing with risks/threats.

The goal of preservation process in ORF is to provide proper 'historical' (older than 1 week) content for actual everyday production and preserve an important part of `company value` and `national culture heritage'. To achieve this goal efficiently, DiMi workflow model was developed for two purposes. First, it can be used as a basis for running “green table” tests prior to installing the actual workflow, which allows finding major risks, potential bottle-necks and other weaknesses in the process. Secondly, DiMi workflow model can help to develop countermeasures and emergency-routines against potential risks/ threats. The purpose of ORF risk management is to minimise loss and damage of valuable content, to produce higher overall output quality, to achieve less deviations from actual output-rate plans and shorter implementation phase within a given time. Negative consequences of something going wrong during digital preservation process can be classified as very high, high and low. Interesting to note that very high risk relates not to the workflow process, but to different issues during preservation process such as budget, relocation of vaults during the process, service provider drop outs etc. The high level of risks related to workflow occurs during material selection and introducing errors on file during transfer and correction/conversion stages.

8.3.3 Identifying risks/threats, negative consequences, controls with cost and time and their analysis

The extract from table provided by ORF to identify risks/threats, their negative consequences for DiMi workflow, any controls in a place with associated costs and times is presented in Table 2. Three measures were asked to be estimated/provided by ORF:

- ***Level of Likeliness*** (LL) for risk/threat to occur, which changes from 1 lowest (not likely) to 5 highest (very likely)
- ***Level of Severity*** (LS) of a given risk/threat, which changes from 1 lowest to 5 highest. The correspondence between LS and percentage of negative consequences: 1 is 5% or less, 2 equals 10%, 3 equals 25%, 4 equals 50%, 5 is up to 100%
- **Estimated Frequency** (EF) is shown in the form: x pH - x times per hour (pD - per day, pW - per week, pM - per month, pY - per Year)

When no information is available about risk/threat, consequences etc, then ***unknown*** is entered. When there is no risks/threats affecting task or their consequences are negligible (dealt on spot for example), then ***n/a*** (not applicable) is entered.

Table 2: Identification of risks/threats in DiMi workflow - extract

Task	Risk/Threat, LL, EF	Causes	Consequences, LS			Control	Cost	Time
			Next Task	Rest of Workflow	Overall			
Content selection based on DiMi criteria	Wrong content selected LL:2, EF: 5 pM	Bad selection criteria	none	None	Depending on affected content high value might get lost, LS:3	Re-training/ Redo Criteria list/ Re-selection / Re-grouping	50 euro per hour	10-60 min

© DAVID consortium: all rights reserved

	Wrong source-material selected LL:2, EF: 5 pM	Wrong decision	None	Additional effort in next steps possible LS : 1	The same, LS:2	The same	50 euro per hour	10-60 min
	Relevant content not selected LL:1, EF: 1 pM	Incorrect information / metadata	None	None	The same, LS:4-5	The same	50 euro per hour	10-60 min
	Too little amount of output LL:1, EF: 2 pM	Logistic causes	Not enough material for order list LS:2	Lagging of the whole process LS:4	none	The same	5000 euro per day	Up to several days
Receive material	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Transfer	Material is damaged LL:1, EF: 2 pM	Fragile material	None	spare-material to be accessed LS:2	restoration efforts LS:3 / Loss LS:5	High service rate / Re-Training	500 euro per Restoration / Loss	2 hours for restoration / 10 hours for Loss
	Material is destroyed LL:1, EF: 1 pY	Pre-damaged material	None	spare-material to be accessed LS:2	Loss LS:5	The same	500 euro per Loss	10 hours for Loss
	Wrong content separation LL:3, EF: 1 pW	Human Error	None	major additional work LS:4	deviations LS:3	The same	50 euro per hour	1 hour
	Wrong IN/OUT LL:2, EF: 2 pM	Wrong information	None	major additional work LS:4	deviations LS:3	The same	50 euro per hour	1 hour

The initial analysis showed that some risks can have very strong effect on the next task, but could have no or very small effect on overall workflow. On the contrary, the other risks practically do not lead to any negative consequences in the next task or remaining workflow, however can cause very series negative consequences for overall preservation process, for example risk: *Relevant content not selected* in Table 2. Having information about level of likeliness, level of severity and estimated frequencies can allow us to use any risks models mentioned above after classification the risks/threats according to SPOT model and discussing scenarios of interest with ORF. In the next section we will show how threats/risks in DiMi workflow can be classified using SPOT model.

8.3.4 Classification of risks/threats according to SPOT model

All risks/threats in DiMi workflow were classified according to SPOT model. Some risks/threats will affect more than one of essential properties of successful digital preservation. Table 3 shows how this classification can be applied to Table 2. If a property is affected by the threat, 1 is put in the table,

© DAVID consortium: all rights reserved

otherwise 0. It is an preliminary attempt and further work is needed in understanding threats/risks in DiMi workflow and SPOT model, and cooperation with ORF. For simplicity, the following notation is used in the table for SPOT model:

- Availability - Av
- Identity - I
- Persistence - P
- Renderability - R
- Understandability - U
- Authenticity - Au
- Neither - N

Table 3: Classification of risks/threats according to SPOT model

Task	Threats/Risks	SPOT						
		Av	I	P	R	U	Au	N
Content selection based on DiMi criteria	Wrong content selected	1	0	0	0	0	0	0
	Wrong source-material selected	1	0	0	0	0	0	0
	Relevant content not selected	1	1	0	0	0	1	0
	Too little amount of output	0	0	0	0	0	0	1
Receive material	n/a	0	0	0	0	0	0	0
Transfer	Material is damaged	1	0	1	0	0	0	1
	Material is destroyed	1	0	1	0	0	0	0
	Wrong content separation	0	0	0	0	0	1	0
	Wrong IN/OUT	0	0	0	0	0	1	0

8.3.5 Further work on Risk management of ORF DiMi workflow

The further work on Risk Management for ORF DiMi workflow consists of the following steps:

- Verification of SPOT model application for DiMi workflow with ORF
- Discuss the scenario of interests and based on these scenarios select
  - the most suitable risk measures
  - ways to display information about risks taken place
- Run scenarios and report results

8.4 Initial Tools for Risk Management

This section describes the initial development work carried out into an analytical tool for risk assessment that uses a simulation-based approach.

8.4.1 Background

The workflow analysis tool is derived from subset of BPMN 2.0. It was created as a risk analysis tool for modelling the workflow, but allowing user input. Signavio was chosen as a suitable starting point, as it has an intuitive web-based graphical user interface supporting the BPMN modelling notation [Signavio].

© DAVID consortium: all rights reserved

jBPM [jBPM] and Signavio are recent tools for carrying out workflow simulations. Signavio is proprietary software which implies no source code to work with and thus is unsuitable to use as a framework. jBPM has hard-coded bpmn2.0 stencil set which makes it impossible to add other risk I/O variable in the model. Therefore it was deemed imprudent to extend on them. However, they both are based on Oryx, which was chosen as a suitable framework for building the risk simulation model.

8.4.2 BPMN 2.0

A generic workflow consists of nodes connected by edges like a graph. An explanation of *nodes* of the archive model derived from BPMN2.0 model is provided below.

Process

Essentially, workflow modelling involves simulating a *process*. Theoretically, a *process* refers the flow of tokens through a sequence of *activities*, initiated from a *start event*, which are finally consumed at an *end event*. A more technical definition of general terms is provided:

Activity: an abstract base class for tasks and sub-processes. It logically represents the type of work to be carried out during a process.

Task: refers to an atomic entity of work to be carried out in the workflow.

Sub-Process: an activity that can be expanded either the given or another workflow, which contains a sequence of tasks being carried out.

Start Event: the entry point of the process. It can generate only one output token.

End event: the finishing point of the process. It encapsulates the result of the process as terminate, error, cancel or none (normal).

Decision Gateway: or exclusive gateway represents a decision point in the process where one path of the flow of tokens diverges into multiple paths. Depending on the Boolean evaluation of a condition, only one path is followed by the flow of the tokens.

Inclusive Gateway: a diverging gateway akin to XOR. But, depending on the Boolean evaluations, enough tokens are generated so that all true evaluating paths are followed.

Parallel Gateway: a diverging gateway, that unlike inclusive or exclusive gateways, does not need any Boolean evaluation. Enough token are generated so that all the paths are followed.

The workflow can be decorated with information about the metadata. These annotations are termed as **Artefacts** and bear no significance on the resulting logical model of the workflow, but provide graphical information to the user.

Collaboration

Another type of workflow modelling involves a *Collaboration*. It involves passing *messages* between different *Participants* to represent exchange of information. A participant may be empty or may contain one process. Every participant is graphically represented as a *pool* and can contain multiple *lane sets*. Each lane-set can contain multiple *lanes*.

The nodes of the process pass the token through edges called *SequenceFlows*. The nodes in a *Collaboration* pass messages through *MessageFlows*.

8.4.3 Archive Model

The BPMN2.0 was altered to custom-fit the requirements of the archive management workflow. Based on the ORF DiMi workflow model on Signavio, it is clearly evident that the archive workflow involves a sequential flow of one token generating from a start event and consumed at an end event. The parallel gateway is the exception where multiple tokens are generated to allow all the paths to be executed.

These restrictions represent the main points of difference between BPMN2.0 model and Archive model. They have been encoded as rules:

- An Activity can have only one input and one output SequenceFlow

© DAVID consortium: all rights reserved

- A Process contains only one StartEvent.
- A Process can contain multiple EndEvents.
- A StartEvent only has one output and none input SequenceFlows.
- An EndEvent only has one input and none output SequenceFlows.
- Any Gateway can only have one input and multiple output SequenceFlows.

8.4.4 Risk Model

- The Archive Risk model involves performers executing tasks in a process that is acting on a file object.
- Each task has associated risks that can damage or corrupt the **file object** of the process.
- **File object:** may represent a collection/package of AV files (digital or physical media) or associated metadata files or a combination of both.
- The corruption is represented by the **SPOT model**.
- The performers are known to the entire Process or Collaboration. They have an error-rate value that determines the number of times that performer will execute a faulty operation that may cause a risk to occur.
- Each risk has associated performers to determine the chain of causation of the faulty operation.
- **Performers** are graphical represented as **Tools** and **Operator** and connected to the Tasks through **Associations**.

8.4.5 Model Class Diagrams

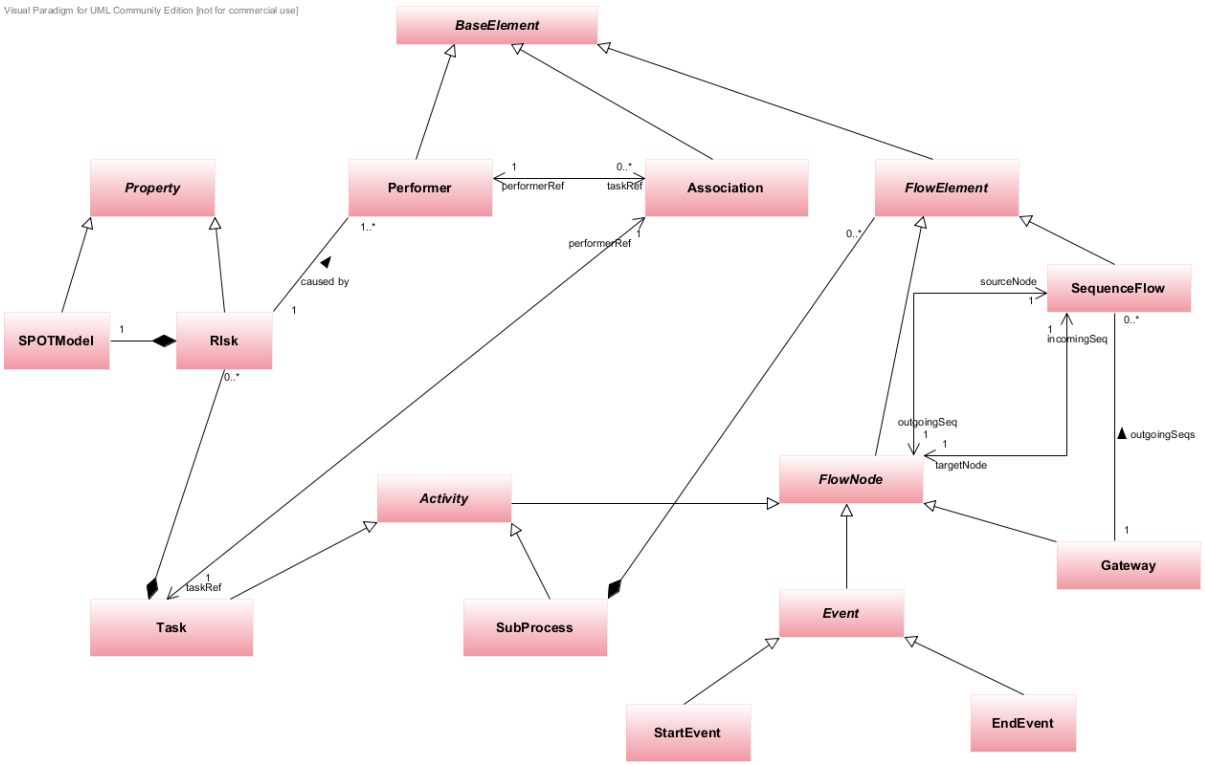

Figure 4: Archive Risk model

BaseElement: abstract base class of the model and all other elements extend it. It represents the model and any element of the model extends it. It defines the namespace of the archive model.

© DAVID consortium: all rights reserved

FlowElement: abstract base class representing that part of the model where the theoretical flow of tokens can take place.

RootElement: abstract base class of the model that represents a simulating entity like a collaboration or process.

Definitions: logical representation of the canvas diagram. It can contain root elements: at most one Collaboration, at least one Process and a list of known Performers on the canvas.

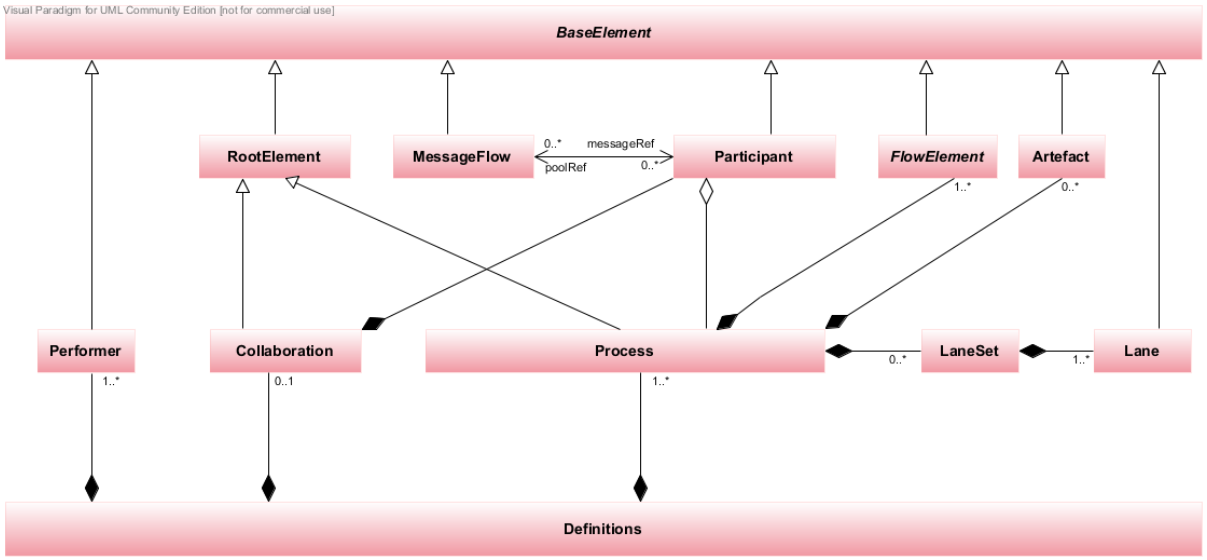

Figure 5: Archive Risk simulation model

8.4.6 Simulation Model

The simulation model only provides a Boolean answer to whether or not the data-file object will become corrupted over a period of time. It does not care about the degree of corruption.

Main steps in the simulating logic:

- Randomly initialise all the known performers as faulty or faultless.
- Beginning at the first task, for each task, determine the list of vulnerable risks.
- Vulnerable risks are those whose set of performers (that may cause the risk to occur) intersect with the set of faulty performers of the task.
- Randomly initialise the vulnerable risks.
- Record the results and repeat the all the steps for the required number of default cycles (derived from user input).

IMPORTANT NOTE: Simulating 100 file objects is equivalent to simulating 1 file object 100 times.

The model determines the individual effect of the risks for each task and each process, on the file object. It does not take into account that propagative effect of the risk in the simulation.

The resulting JSON has this form:

```
A process json object  :
{
"Task":[{
        "name:"string",
        "performer":[{
                        "id":"string",
                        "faulty":"false"
                },

                {...
                        //case #1: assuming all performers are faultless
```

```
                    }]
            },
            {
             "name":"string",
             "performer":[{
                             "id":"string",
                            "faulty":"true"
                        },
                        {...
                            //case #2: at least one performer is faulty
                        }]
              "risk":[{
                        "id":"string",
                        "name":"string",
                        "occurred":"false" //case #3: risk did not occur
                       },
                       {
                        "id":"string",
                        "name":"string",
                        occurred":"true" //case #4: risk did occur and control applied
                        "control":"true"
                       },
                       {
                        "id":"string",
                        "name":"string",
                        "occurred":"true"  //case  #5:  risk  did  occur  and  control  not
applied
                        "control":"false",
                        "spot":{
                                "authenticity:"boolean",
                                "availability":"boolean",
                                     .
                                     .
                                     .
                                     .
                               }
                        }]
                }]
            }
```

8.4.7 Further Work to be done

- Parallel gateway needs to be added to the simulation logic
- The results can be accumulated and sent back to the server to be displayed in suitable graphical format.
- The propagative logic of risk simulation can be added.

8.5 Preservation metadata model

This section discusses the representation of metadata of preservation processes, which supports both risk management tools and adaptive (e.g. rule-based) preservation workflows. The model described here complements the set of technical metadata properties stored with the content by documenting the processing applied.

© DAVID consortium: all rights reserved

8.5.1 Scope

As part of the preservation metadata of an audio-visual content, the scope of the preservation process metadata model is to document the history of creation and processing steps applied, as well as their parameters.

The model represents the preservation actions that were actually applied, i.e., a linear sequence of activities, with the option to have a hierarchy for grouping activities.

The model supports a set of specific types of activities in the model (e.g., digitisation, with possible further specialisations, e.g. film scan), in order to improve interoperability between preservation systems.

The model also describes the parameters of these activities, beyond a generic key/value structure. There should be a core set of well-defined properties, with type, and storing the value used when processing the item described. Of course, in addition there can be a key/value structure for supporting extensions, but a small set of core properties is be defined for an activity.

A specific set of these parameters are the description of tools/devices used in these processes, as well as their parameters.

8.5.2 Model definition

The model is designed around three main groups of entities: content entities (DigitalItems, their Components and related Resources), Activities and Operators (Agent, Tool) and their properties. The content entities are created, used or modified in an Activity, which involves Operators that contribute to performing the Activity. The basic entities of the model and their relations are shown in Figure 6, and described in Table 4.

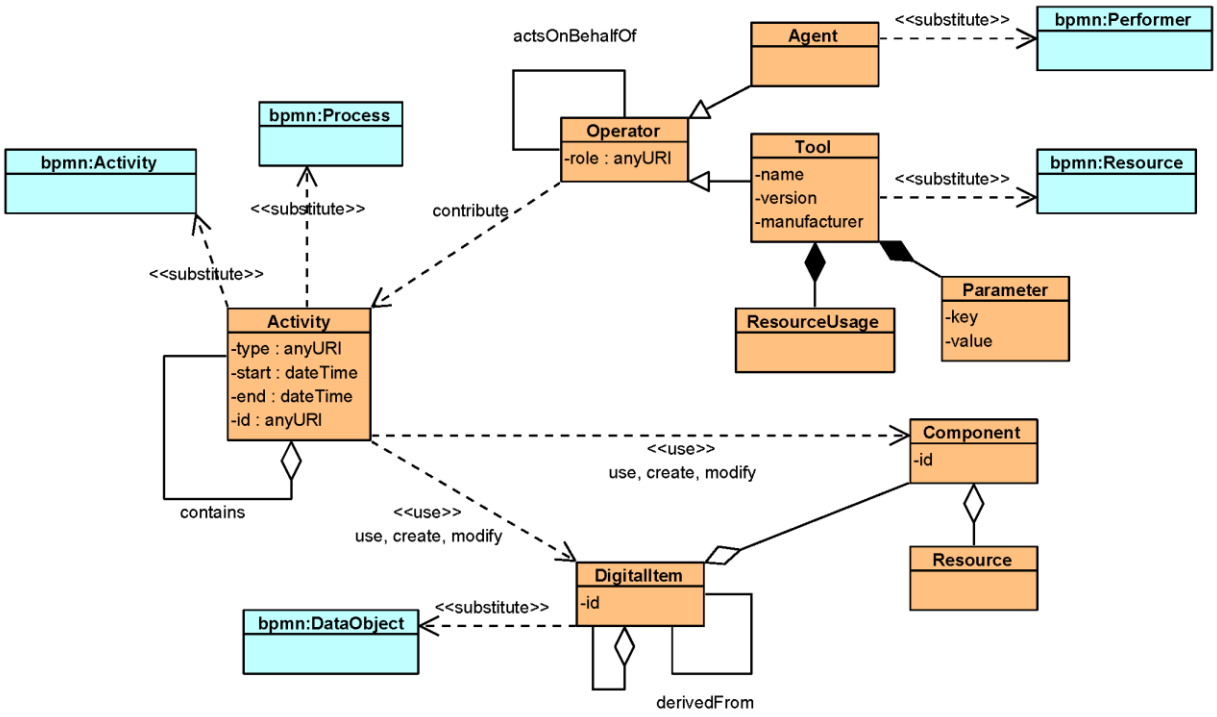

Figure 6: Entities of the preservation data model, their relations and the most important core properties. Blue entities are related BPMN entities.

Entity	Description	Relations
Activity	An action in the lifecycle of the content item	*contains* Activity, i.e. is composed of other, more fine-grained Activities *uses* a DigitalItem or a Component, this relation is further distinguished into *uses, creates, modifies*

© DAVID consortium: all rights reserved

Entity	Description	Relations
DigitalItem	An intellectual/editorial entity to be preserved, a representation of such an entity or an essence.	*Aggregates* other DigitalItems (e.g., the representations of an intellectual/editorial entity, the essences constituting the representation) *Aggregates* Components (e.g., the bitstreams of an essence) *isDerivedFrom* other DigitalItems (e.g., by migration)
Component	A component is the binding of a resource to a set of metadata. A component itself is not an item; components are building blocks of items.	*Aggregates* Resources
Resource	A resource is an individually identifiable content file or bitstream such as a video or audio clip, an image, or a textual asset. A resource may also potentially be a physical object. All resources shall be locatable via an unambiguous address.	
Operator	An entity contributing to the completion of an Activity by performing (part of) it or being used to perform it.	*Performs* an Activity, the type of involvement is further specified by the Operator's role attribute *Composition* of Parameters and ResourceUsage information *actsOnBehalfOf* another Operator
Agent	A person or organisation involved in performing an activity.	
Tool	A device or software involved in performing an activity.	
Parameter	A key/value structure for holding information about Operators.	
ResourceUsage	A structure holding information about the resource usage by the Operator when performing the activity.	

Table 4: Description of the model entities.

Activities have start and end times, and their inputs/outputs are identified. This enables the reconstruction of the execution order and dependencies, without an explicit description of serial or parallel activities, and without having specific start/end events. Having a generic activity and no discrimination into tasks and sub-processes harmonises handling preservation process descriptions with different granularity.

Types of activities are modelled by reference to a controlled vocabulary, rather than defining the classes in the model (see below).

Subclasses of DigitalItem (such as supported in MPEG-21, PREMIS) can be optionally added, but are not needed for the purpose of describing preservation history. However, the levels of component/resources (DigitalItem has Components has Resources, also found in other models such as MPEG-21) has been added, as it allows describing activities working on components. This distinction also allows describing DigitalItems and components without related resources, which is useful for describing preservation activities that failed, and thus lack the essence in the package, but should be kept to support risk assessment.

© DAVID consortium: all rights reserved

Essence and metadata can only be reliably identified if they lie in the same package (SIP/AIP/DIP according to OAIS [OAIS] terminology), external data can be referenced (preferably with a URI). Any metadata is represented in the context of a DigitalItem, which is used, created or modified by activities.

Parameters and Resource Usage have been added as separate classes. Optionally, instances of these classes can be used to complement the parameters of Tool. If needed, specific subclasses of Tool can be defined with additional required parameters.

8.5.3 Relation to BPMN

BPMN [BPMN] is a good choice for representing processes of a preservation workflow, to configure, simulate and execute them. BPMN does not fully meet the requirements for the information we want to represent as part of the presentation metadata, thus extensions are needed at some points. On the other hand, these metadata document what has actually been done, thus many features of BPMN, such as gateways, events, looping etc., are not needed by the model.

From our experience, there are quite severe incompatibilities between different implementations using BPMN. The recent establishment of the model interchange working group[1] shows that this is indeed an issue. A second aspect that adds to complexity is the duality of the BPMN standard, i.e., that a BPMN document contains both a process definition and a diagram interchange description. Both issues are problematic when considering this as a format to be used in long-term preservation.

In this document, BPMN refers to Business Process Model and Notation (BPMN), Version 2.0, January 2011.

Our model makes a number of simplifications over BPMN, in order to eliminate constructs not needed based on the requirements described above, and in order to facilitate interoperability. However, interoperability with BPMN is provided. The core entities can be aligned with BPMN, mandatory BPMN attributes can be added as optional ones. This allows implementing conversion from/to BPMN as an XSL transform. Conversion from BPMN to generate placeholders for the activities to be run in a process is considered the more common case. The inverse conversion would only be needed for generating processes that rerun a chain of preservation activities executed previously, e.g. to recreate an item from an earlier generation that cannot otherwise be recovered.

Because of the overhead in inheritance structure and type definitions in BPMN, direct import of the schema does not seem useful, but a simple1:1 conversion (mostly changing namespaces, stripping unsupported attributes) is feasible, indicated by the substitution relations in Figure 6.

For the model, a generic activity was found to be sufficient. Mapping rules to other BPMN constructs can be defined as follows:

- A process corresponds to an activity without parent activity.
- A task corresponds to an activity without child activities.
- A sub-process corresponds to an activity with child activities.

The mapping of users and tools/devices can be done as follows:

- BPMN performer is defined as a Agent, linked to an Activity via a resource role
- BPMN resource is linked to an activity via a resource role, and provides list of parameters indirectly via ParameterBinding
- BPMN participant is only linked to process and orchestration, not to an activity
- It does not seem necessary to replicate this structure, but it could be converted back/forth to from the proposed representation.

8.5.4 Relation to other preservation data models

The model has been designed considering interoperability with models for representing preservation metadata and provenance from beyond the audio-visual domain, in particular with PREMIS [PREMIS] and the W3C Provenance data model [PROV-DM, 2013].

In PREMIS, there are bidirectional relationships between Object and Event, Event and Agent, relationships between Objects. Object maps to DigitalItem, Event to Activity and Agent to Operator.

[1] http://www.omgwiki.org/bpmn-miwg/doku.php

© DAVID consortium: all rights reserved

These relationships can be represented with the existing model, with inverting the type of relation if needed. Rights is an additional core entity in PREMIS, which is considered out of scope of this model.

The PROV-DM uses one more GenericEntity, which maps to DigitalItem, Agent maps to Operator and Activity is equivalent in both models. The model includes the following relations, which can also be mapped to the proposed model (inverting the relation in some cases):

- Entity wasGeneratedBy Activity
- Activity used Entity
- Entity wasDerivedFrom Entity
- Activity wasAssociatedWith Agent
- Agent actedOnBehalfOf Agent

The relation Entity wasAttributedTo Agent, is neither supported by our model nor PREMIS. The other considered initiative was the work in progress by MPEG to define the Multimedia Preservation Application Format (MP-AF). At the time of writing, the 3rd Working Draft of MP-AF is available [MP-AF], which is compatible with the proposed model as a result of the inputs from the DAVID project to this MPEG group.

8.5.5 Specific activities

The following hierarchy of activities is derived from D2.3 process descriptions, iModel and PrestoPRIME documentation.

Activities

- Acquisition/Recording
- Ingest
- Migration/Distribution copy generation/proxy generation
    - Analogue transfer recording
    - Digitisation
        - Scanning
        - Transfer Recording
    - Digital Migration
        - Transwrapping
        - Transcoding
- Cleaning
- Checking
    - Checksum generation
    - Checksum verification
    - Integrity checking
    - Quality Control
        - CRT-based
        - File-based
        - Verification
            - Technical
            - editorial
- Material selection
- Material handling
    - Taking
    - Shelving
    - Shipping
- Mapping control
- Preview control

© DAVID consortium: all rights reserved

- Carrier liquidation
- Metadata
  - Enrichment
  - Modification
- Repair, Correction
  - Replace copy
  - Local repair
  - Restoration
- Access
  - Delivery

8.5.6 Properties

Activities have identifier, type, start and end date/time as their properties. Some properties of the activity, that are evident from the related DigitalItems (e.g. source/target format) or from the properties of related tool description, are not specified as properties of the activity.

At least one Operator must be associated with an activity. For all activities that require the use of devices or software tools, those must be specified.

Further specific properties are currently being defined.

8.5.7 Tools

All tools have a name, version and manufacturer as general properties.

Tool properties	**Properties**

Acquisition/Recording
• Camera
  ○ Film
  ○ Analogue
  ○ Digital
• Tape recorder
  ○ Analogue
  ○ Digital
• Optical disk recorder
• Digital capture board
• Film scanner
• Telecine
	EBU Tech 3349 (HIPS-META) might contain basic metadata (to be validated, if this is sufficient)

Checksum generator/verifier	Type of checksum, granularity
Quality Control	EBU QC checks and parameters

Table 5: Additional properties for specific tool types.

© DAVID consortium: all rights reserved

9 Conclusions

This report has defined the scope of digital damage for the DAVID project and has motivated the need for approaches to long-term quality assurance of digital assets.

Best practice has developed progressively in the archive domain, both as a result of the wealth of experience within the community on preservation of analogue artefacts, but also with the help of research projects to apply these lessons to the new challenges of digital preservation. The infrastructure for digital archives is IT-based, which enables new approaches to be applied to the problem of preservation, while others approaches can be modified from previous experience. However, the effectiveness of strategies for preventing, reducing and recovering from loss is not clear. The approaches represent a 'bag of tools', which are typically applied using a 'defence in depth' approach without clearly articulating the cost/benefit of each layer of mitigation.

Experientially, thus far, the expert design of preservation systems has proven mostly successful. IT-based systems of sufficient integrity, coupled with data replication, have proven effective at keeping the level of digital damage to a minimum. In the rare cases where damage has been experienced, loss of the asset has been avoided through recovery from a replica copy. This is certainly the experience of the archive partners in the DAVID project, INA and ORF, and (natural disasters notwithstanding) appears to be true in the wider community.

The principal problem experienced seems to be one of incompatibility, i.e. the choice of codecs/wrappers (or specific implementations of these) that pose a problem many years after the generation/migration of the digital file. The problem to solve then becomes one of format selection (possibly simplifying the storage of the data in basic streams of essence), profile specification and provenance tracking (e.g. recording the metadata about the options used when generating the new files).

Prototype preservation planning tools have been developed as part of recent research and development efforts in the community. Such tools have helped to promote understanding of the relationship between cost and risk in preservation systems. However, these tools need to be extended with an appreciation of the preservation workflows, such that task of risk management can be combined with business process design and, eventually, execution. The focus on business process risk management arises directly from the observation that specific and relevant problems in the preservation community are rarely due to random failure (such as corruption events), but to systematic errors, such as format choices, tool misconfiguration and process changes.

Within the DAVID project, WP3 focuses its efforts on techniques and tools for metadata capture and risk management decisions that affect format incompatibility and other systematic failures, rather than on corruption arising from random failure. The aim is to bring these tools together to form risk management framework that allows archive specialists to articulate risks during the specification of business processes, and to use simulation-based approaches to analyse the effects of deploying risk treatment strategies. Preservation metadata will provide evidence to support and improve the risk assessment approach.

© DAVID consortium: all rights reserved

10 References

[Addis, 2010] M. Addis, M. Jacyno, M. Hall-May, and R. Wright, Storage Strategy Tools. International Association of Sound and Audiovisual Archives Journal, no. 38, Jan 2012.

[Addis, 2013] M. Addis, 8k traffic jam ahead, PrestoCentre blog, Apr. 2013. https://www.prestocentre.org/blog/8k-traffic-jam-ahead

[AVArtifactAtlas] A/V Artifact Atlas, Bay Area Video Coalition. http://preservation.bavc.org/artifactatlas/index.php/A/V_Artifact_Atlas

[AVPreserve] AudioVisual Preservation Solutions. http://www.avpreserve.com/

[Avid, 2006] MXF Unwrapped, Avid Post Production, 2006. http://www.avid.com/static/resources/common/documents/mxf.pdf

[Bai, 2013] X. Bai, R. Krishnan, R. Padman, H.J,Wang. On Risk Management with Information Flows in Business Processes. Information Systems Research, 24(3):731-749, Nov. 2013.

[Becker, 2010] C. Becker, H. Kulovits, and A. Rauber,Trustworthy Preservation Planning with Plato, ERCIM News 80, p.p. 24–25, Jan. 2010. http://ercim-news.ercim.eu/images/stories/EN80/EN80-web.pdf

[Berger, 1980] J. Berger. Statistical Decision Theory: Foundations, Concepts, and Methods. Springer-Verlag, New York, 1980.

[Besser, 2000] H. Besser, Digital longevity, Chapter in M. Sitts (ed.) Handbook for Digital Projects: A Management Tool for Preservation and Access, Andover MA: Northeast Document Conservation Center, 2000, pp. 155-166. http://besser.tsoa.nyu.edu/howard/Papers/replaced/sfs-longevity.html

[Bilgin, 2003] A. Bilgin, Z. Wu, and M. W. Marcellin, Decompression Of Corrupt JPEG2000 Codestreams, In Proceedings of the Data Compression Conference, 2003.

[BPMN] Object Management Group Business Process Model and Notation. http://www.bpmn.org/

[Brown, 2008] A. Brown, Digital Preservation Guidance Note 1: Selecting file formats for long-term preservation, The National Archives, Aug. 2008. http://www.nationalarchives.gov.uk/documents/selecting-file-formats.pdf

[Buonara, 2008] P. Buonara and F. Liberati, A Format for Digital Preservation of Images - A Study on JPEG 2000 File Robustness, D-Lib Magazine, Vol. 14, No. 7/8, Jul. 2008.

[CEL] MPEG-21 Contract Expression Language. http://mpeg.chiariglione.org/standards/mpeg-21/contract-expression-language

[Chayka, 2012] K. Chayka, Hurricane Sandy Highlights the Problems of Digital Archives, Hyperallergic article, Nov. 2012. http://hyperallergic.com/60598/eyebeam-hurricane-sandy-flooding/

[Chivers, 2012a] L. Chivers, Truncated JPEG2000, OPF Knowledge Base Wiki. http://wiki.opf-labs.org/display/REQ/Truncated+JPEG2000

[Chivers, 2012b] L. Chivers, Shifted Crop Corruption, OPF Knowledge Base Wiki. http://wiki.opf-labs.org/display/REQ/Shifted+Crop+Corruption

[Cochran, 2012] E. Cochran, Rendering Matters - Report on the results of research into digital object rendering, Archives New Zealand, Jan. 2012. http://archives.govt.nz/sites/default/files/Rendering_Matters.pdf

[Comité des Sages, 2011] E. Niggemann, J. de Decker, M. Lévy, The New Renaissance, Report of the Comité des Sages, Jan. 2011.

© DAVID consortium: all rights reserved

http://ec.europa.eu/information_society/activities/digital_libraries/doc/refgroup/final_report_cds.pdf

[Conforti, 2011] R. Conforti, G. Fortino, M. La Rosa, A. H. M. ter Hofstede, History-Aware, Real-Time Risk Detection in Business Processes, On the Move to Meaningful Internet Systems, LNCS Volume 7044, pp. 100-118, 2011.

[CPDP] Cylinder Preservation and Digitization Project, Department of Special Collections, Donald C. Davidson Library, University of California, Santa Barbara. http://cylinders.library.ucsb.edu/

[Cunningham, 2007] S. Cunningham and P. de Nier, File-based Production: Making It Work In Practice, BBC Research White Paper, WHP 155, Sep. 2007. http://downloads.bbc.co.uk/rd/pubs/whp/whp-pdf-files/WHP155.pdf

[DigitalFormats] Profile 4 for JPEG 2000, Part 1, Core Coding, Sustainability of Digital Formats Planning for Library of Congress Collections. http://www.digitalpreservation.gov/formats/fdd/fdd000213.shtml

[DSpace] http://www.dspace.org/

[Duffle, 1997] An overview of value at risk. J. Derivatives, 4(3), pp. 7-49

[Gledson, 2010] A. Gledson and P. Watry, Media formats, identification methods and implementations for multivalent preservation, PrestoPRIME Internal Deliverable ID3.3.1, May 2010. https://prestoprimews.ina.fr/public/deliverables/PP_WP3_ID3.3.1_multivalent_R0_v1.02.pdf

[Graf, 2013] R. Graf and S. Gordea, A Risk Analysis of File Formats for Preservation Planning, In proceedings of 10th International Conference on Preservation of Digital Objects, Sep. 2013.

[Green, 2003] D. L. Green (chair), et al, The NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials, v1.1, ch. XIII, National Initiative for a Networked Cultural Heritage, 2003. http://www.ninch.org/guide.pdf

[Gattuso, 2013] J. Gattuso, Exploring the impact of Bit Rot, National Library of New Zealand, Feb. 2013. http://www.openplanetsfoundation.org/system/files/Bit%20Rot_OPF_0.pdf

[Heydegger, 2008] V. Heydegger, Analysing the impact of file formats on data integrity, Archiving Conference, Society for Imaging Science and Technology, 2008.

[Heydegger, 2009] V. Heydegger, Just One Bit in a Million: On the Effects of Data Corruption in Files, Research and Advanced Technology for Digital Libraries, LNCS Volume 5714, 2009, pp. 315-326.

[iModel] iModel v1.0 documentation, Workflows. http://prestoprime.it-innovation.soton.ac.uk/imodel/docs/workflows.html

[ISO16363, 2012] ISO 16363:2012, Space data and information transfer systems - Audit and certification of trustworthy digital repositories.

[ISO31000, 2009] ISO 31000:2009, Risk management - Principles and guidelines. http://www.iso.org/iso/home/standards/iso31000.htm

[jBPM] http://www.jboss.org/jbpm

[Kaufman, 2013] P. B. Kaufman, Assessing the Audiovisual Archive Market - Models and Approaches for Audiovisual Content Exploitation, PrestoCentre white paper, 2013. https://www.prestocentre.org/library/resources/assessing-audiovisual-archive-market

[Kula, 2002] S. Kula, Appraising Moving Images: Assessing the Archival and Monetary Value of Film and Video Records, Lanham, Maryland and Oxford: Scarecrow Press, 2000.

[Lacinak, 2010] C. Lacinak, A Primer on Codecs for Moving Image and Sound Archives & 10 Recommendations for Codec Selection and Management, Audiovisual Preservation

© DAVID consortium: all rights reserved

Solutions, 2010. http://www.avpreserve.com/wp-content/uploads/2010/04/AVPS_Codec_Primer.pdf

[Lavoie, 2012] B. Lavoie, Preservation metadata as an evidence base for risk assessment, iPRES, Oct. 2012. http://www.loc.gov/standards/premis/pif-presentations-2012/phc_brian.pdf

[Lawrence, 2000] G. W. Lawrence, W. R. Kehoe, O. Y. Rieger, W. H. Walters, A. R. Kenney, Risk Management of Digital Information: A File Format Investigation, Council on Library and Information Resources, 2000. http://www.clir.org/pubs/reports/pub93/reports/pub93/pub93.pdf

[LeFurgy, 2012] B. LeFurgy, Bits Breaking Bad: The Atlas of Digital Damages, The Signal, Oct. 2012. http://blogs.loc.gov/digitalpreservation/2012/10/bits-breaking-bad-the-atlas-of-digital-damages/

[LeFurgy, 2013] B. LeFurgy, Is JPEG-2000 a Preservation Risk? The Signal, Jan. 2013. http://blogs.loc.gov/digitalpreservation/2013/01/is-jpeg-2000-a-preservation-risk/

[LoC] Sustainability Factors, Sustainability of Digital Formats Planning for Library of Congress Collections. http://www.digitalpreservation.gov/formats/sustain/sustain.shtml

[LOCKSS] Lots Of Copies Keeps Stuff Safe, http://www.lockss.org/

[van Malssen, 2013] K. van Malssen, When the 'Worst' Happens: How a disaster can change our perspective on the motivations and priorities for digital AV preservation, Screening the Future, 2013.

[MCO] MPEG-21 Media Contract Ontology. http://mpeg.chiariglione.org/standards/mpeg-21/media-contract-ontology

[MP-AF] Multimedia Preservation Application Format. http://mpeg.chiariglione.org/standards/mpeg-a/multimedia-preservation-application-format

[MXF_OP1a_JP2_LL] MXF File, OP1a, Lossy JPEG 2000 in Generic Container, Sustainability of Digital Formats, Planning for Library of Congress Collections. http://www.digitalpreservation.gov/formats/fdd/fdd000206.shtml

[MXF_OP1a_JP2_LSY] MXF File, OP1a, Lossless JPEG 2000 in Generic Container, Sustainability of Digital Formats, Planning for Library of Congress Collections. http://www.digitalpreservation.gov/formats/fdd/fdd000206.shtml

[OAIS] Reference Model for an Open Archival Information System (OAIS), Recommended Practice, issue 2, Consultative Committee for Space Data Systems, Jun. 2012. http://public.ccsds.org/publications/archive/650x0m2.pdf

[OpenAXF, 2011] Archive eXchange Format white paper, Front Porch Digital, 2011. http://www.openaxf.org/pdf/AXF%20White%20Paper.pdf

[Panzer-Steindel, 2007] B. Panzer-Steindel, Data integrity CERN/IT, Draft 1.3, Apr. 2007. http://indico.cern.ch/getFile.py/access?contribId=3&sessionId=0&resId=1&materialId=paper&confId=13797

[Petersen, 2007] M. K. Petersen, T10 Data Integrity Feature (Logical Block Guarding), Linux Storage & Filesystem Workshop, Feb. 2007.

[Prabhakaran, 2005] V. Prabhakaran , L. N. Bairavasundaram , N. Agrawal, H. S. Gunawi , A. C. Arpaci-dusseau , R. H. Arpaci-dusseau, IRON file systems, In Proceedings of the 20th ACM Symposium on Operating Systems Principles, 2005.

[PREMIS] PREMIS Data Dictionary for Preservation Metadata, Library of Congress Standard. http://www.loc.gov/standards/premis/

[PrestoPRIME] EC FP7 231161 PrestoPRIME. http://www.prestoprime.org/

[PROV-DM, 2013] PROV-DM: The PROV Data Model, W3C Recommendation, Apr. 2013. http://www.w3.org/TR/prov-dm/

© DAVID consortium: all rights reserved

[Reason, 2000] J. Reason, Human error: models and management, British Medical Journal 320 (7237): 768–77. http://www.bmj.com/content/320/7237/768

[REWIND] EC FP7 268478 REWIND, Reverse Engineering of Audio-visual Content Data. http://www.rewindproject.eu/

[Rockafellar, 2000] R.T. Rockafellar, A. Uryasev. Optimization of conditional value at risk. J. Risk, 2(3), PP. 21-42.

[Rosemann, 2005] M. Rosemann, M. zur Muehlen, Integrating Risks in Business Process Models, ACIS Proceedings, 2005. http://aisel.aisnet.org/acis2005/50/

[Rosenthal, 2007] D. Rosenthal, Format Obsolescence: the Prostate Cancer of Preservation, DSHR's Blog, May 2007. http://blog.dshr.org/2007/05/format-obsolescence-prostate-cancer-of.html

[Rosenthal, 2009a] D. Rosenthal, Postel's Law, DSHR's Blog, Jan. 2009. http://blog.dshr.org/2009/01/postels-law.html

[Rosenthal, 2009b] D. Rosenthal, Are format specifications important for preservation? Jan. 2009. http://blog.dshr.org/2009/01/are-format-specifications-important-for.html

[Rosenthal, 2013] D. Rosenthal, D. L. Vargas, Distributed Digital Preservation in the Cloud, International Journal of Digital Curation, Vol. 8, No. 1, 2013. http://www.ijdc.net/index.php/ijdc/article/view/8.1.107

[Rouse] M. Rouse, JBOD (just a bunch of disks or just a bunch of drives), SearchStorage TechTarget. http://searchstorage.techtarget.com/definition/JBOD

[Sarykalin, 2008] S. Sarykalin, G. Serraiono, S. Uryasev. Value-at-Risk vs Conditional Value-at-Risk in Risk Management and Optimisation. Tutorials in Operations Research, 2008

[Sienou, 2007] A. Sienou, E. Lamine, A. Karduck, H. Pingaud, Conceptual Model of Risk: Towards a Risk Modelling Language, Web Information Systems Engineering, LNCS 4832, pp 118-129, 2007.

[Signavio] http://www.signavio.com/

[Sitts, 2000] M. K. Sitts, ed., Handbook for Digital Projects: A Management Tool for Preservation and Access, Northeast Document Conservation Center, Andover, Mass., 2000. http://www.nedcc.org/assets/media/documents/dman.pdf

[ST2034-1] TC-31FS WG-30 Archive eXchange Format (AXF) Part 1, Society of Motion Picture & Television Engineers, Oct. 2013. https://kws.smpte.org/kws/public/projects/project/details?project_id=93

[Sumanta] B. K. Samanta, File-based QC – Delivering Content with Confidence, Interra Systems. http://www.interrasystems.com/pdf/WP_File-based%20QC%20Delivering%20Content%20with%20Confidence.pdf

[Sun, 2010] Solaris ZFS Administration Guide, ch. 11, ZFS Troubleshooting and Pool Recovery, Sun Microsystems, 2010. http://docs.oracle.com/cd/E19082-01/817-2271/817-2271.pdf

[Suriadi, 2012] S. Suriadi, B. Weiß, et al, Current Research in Risk-Aware Business Process Management - Overview, Comparison, and Gap Analysis, BPM Center Report BPM-12-13, 2012. http://bpmcenter.org/wp-content/uploads/reports/2012/BPM-12-13.pdf

[Varra, 2012] J. Varra, Selecting a Digital File Format for France's Professional Television Archive, SMPTE Mot. Imag. J, 121:(1) 51-5, Jan./Feb. 2012.

[Vermaaten, 2012] S. Vermaaten, B. Lavoie and Pr. Caplan, Identifying Threats to Successful Digital Preservation: the SPOT Model for Risk Assessment, D-Lib Magazine vol. 18, no. 9/10, Sep./Oct. 2012. http://www.dlib.org/dlib/september12/vermaaten/09vermaaten.html

[VideoHelp, 2008] "Video corruption when trying to convert MPEG-2 to DVD-compliant", VideoHelp.com forum post, Jan 2008. http://forum.videohelp.com/threads/284080-Video-corruption-when-trying-to-convert-MPEG-2-to-DVD-compliant

© DAVID consortium: all rights reserved

[VidiCert] VidiCert, Joanneum Research. http://www.joanneum.at/en/digital/products-solutions/vidicert.html

[de Vries, 2013] J. de Vries, D. Schellenberg, L. Abelmann, A. Manz and M. Elwenspoek, Towards Gigayear Storage Using a Silicon-Nitride/Tungsten Based Medium, arXiv preprint arXiv:1310.2961, 2013. http://arxiv.org/pdf/1310.2961v1.pdf

[Weatherspoon, 2002] H. Weatherspoon and J. D. Kubiatowicz, Erasure Coding vs. Replication: A Quantitative Comparison, Peer-to-Peer Systems, LNCS Volume 2429, 2002, pp 328-337.

[Weerakkody] R. Weerakkody, Multiple Sub Stream Error Resilient Video Coding for Audio Visual Archiving Applications, BBC (R&D). http://downloads.bbc.co.uk/rd/projects/avatar_m/documents/FinalManuscript_721-040.pdf

[van der Werf] T. van der Werf, Preservation Health Check: introduction to the pilot, iPRES, Oct. 2012. http://www.loc.gov/standards/premis/pif-presentations-2012/phc_titia.pdf

[Wheatley, 2011] P. Wheatley, Unknown JPEG2000 characteristics presents risks to quality, preservation and access, OPF Knowledge Base Wiki. http://wiki.opf-labs.org/display/AQuA/Unknown+JPEG2000+characteristics+presents+risks+to+quality%2C+preservation+and+access

[Wheatley, 2012] P. Wheatley , JISC1 19th Century Digitised Newspapers (BL), OPF Knowledge Base Wiki. http://wiki.opf-labs.org/display/AQuA/JISC1+19th+Century+Digitised+Newspapers+%28BL%29

[Wheatley, 2013] P. Wheatley, Digital Preservation and Data Curation Requirements and Solutions, Open Planets Foundation Knowledge Base Wiki. http://wiki.opf-labs.org/display/REQ/Digital+Preservation+and+Data+Curation+Requirements+and+Solutions

[XCL] eXtensible Characterisation Language. http://planetarium.hki.uni-koeln.de/planets_cms/index.php

[Zhang, 2013] J. Zhang, M. Gecevičius, M. Beresna, P. G. Kazansky, 5D Data Storage by Ultrafast Laser Nanostructuring in Glass, Conference on Lasers and Electro-Optics, 2013. http://www.orc.soton.ac.uk/fileadmin/downloads/5D_Data_Storage_by_Ultrafast_Laser_Nanostructuring_in_Glass.pdf

© DAVID consortium: all rights reserved

11 Glossary

Terms used within the DAVID project, sorted alphabetically.

AIP	Archive Information Package
AV	Audio-Visual
AXF	Archive eXchange Format
BPM	Business Process Model
BPMN	Business Process Modelling Notation
COI	Cost Of Inaction
CRC	Cyclic Redundancy Check
CVaR	Conditional Value at Risk
DAVID	Digital AV Media Damage Prevention and Repair
DIP	Dissemination Information Package
FFV1	FFmpeg video codec 1
GOP	Group Of Pictures
HD	High Definition
HSM	Hierarchical Storage Management
IT	Information Technology
JBOD	Just a Bunch Of Disks
KAG	KLV Alignment Grid
KLV	Key Length Value
LOCKSS	Lots Of Copies Keeps Stuff Safe
LTO	Linear Tape Open
MD5	Message Digest function 5
MPEG	Moving Picture Expert Group
MTTF	Mean Time To Failure
MXF	Material eXchange Format
OAIS	Open Archival Information System
PREMIS	Preservation Metadata: Implementation Strategies
OP	Operational Profile
QA	Quality Assurance
QC	Quality Check
RAID	Redundant Array of Independent Disks
ROI	Return On Investment
SD	Standard Definition
SHA-1	Secure Hash Algorithm 1
SIP	Submission Information Package
SMPTE	Society of Motion Picture & Television Engineers
SPOT	Simple Property-Oriented Threat (model)

© DAVID consortium: all rights reserved

URI	Uniform Resource Identifier
VaR	Value at Risk

Partner Acronyms

CTI	Cube-Tec International GmbH, GE
HSA	HS-ART Digital Service GmbH, AT
INA	Institut National de l'Audiovisuel, FR
ITInnov	University of Southampton - IT Innovation Centre, UK
JRS	JOANNEUM RESEARCH Forschungsgesellschaft mbH, AT
ORF	Österreichischer Rundfunk, AT

Acknowledgement: The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 600827.

© DAVID consortium: all rights reserved