DIGITAL AV MEDIA DAMAGE
PREVENTION AND REPAIR

Final IT-based Tools and Strategies for Avoiding, Mitigating and Recovering from Digital AV Loss

&

Final Conceptual Risk Management Framework and Tools for Digital AV Preservation

SEVENTH FRAMEWORK PROGRAMME

Deliverable D3.3

DAVID identifier: DAVID-D3.3-Final-IT-Strategies-and-Risk-Framework-v1.0.docx

Deliverable number: D3.3

Author(s) and company: Vegard Engen (ITInnov), Galina Veres (ITInnov), Martin Hall-May (ITInnov), Jean-Hugues Chenot (INA), Christoph Bauer (ORF), Werner Bailer (JRS), Martin Höffernig (JRS), Jörg Houpert (CTI)

Abstract: This is the final report from the DAVID project on IT-based tools, strategies and risk management for digital Audio-Visual (AV) preservation.

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. The problems affecting digital AV content and the strategies against AV loss adopted by INA and ORF are discussed in this report.

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. This report presents a conceptual risk management framework responding to the requirements of archive risk management, and presents a set of tools developed by the DAVID partners. This report also contains a specific use case of a real-life MXF file repair process at ORF, showing how risk modelling and simulation can be used in practice.

Internal reviewers: Christoph Bauer (ORF)

Work package / task: WP3 T3.1 & T3.2

Document status: Final

Confidentiality: Public

DOI: 10.7800/304DAVID33

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.

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 vi

3 List of Tables vii

4 Executive Summary 8

5 Introduction 10

5.1 Purpose of this Document 10

5.2 Scope of this Document 10

5.3 Status of this Document 10

5.4 Related Documents 11

6 Problems Affecting Digital AV Contents 12

6.1 Digital versus Analogue Audio-visual Content 12

6.2 What is Digital Damage? 12

6.3 Preservation Metadata for Digital AV Content 16

7 IT-based Strategies Against Digital AV Loss 18

7.1 Avoiding, Mitigating and Recovering from Digital AV Loss 18

7.1.1 Robustness of Technology 18

7.1.2 Robustness of People 23

7.1.3 Robustness of Processes 24

7.2 IT-based Preservation at INA 25

7.2.1 Physical Security of AV Data 25

7.2.2 Backups of AV Data 26

7.2.3 System Security of AV Data 26

7.2.4 AV Data Security Overall in INA 26

7.2.5 Comparison with Analogue Workflows 27

7.3 IT-based Preservation at ORF 27

7.3.1 Physical Security of AV Data 27

7.3.2 Backups of AV Data 27

7.3.3 Security of AV Data and File Conformance 27

7.3.4 Comparison with Analogue Workflows 28

8 Conceptual Risk Management Framework 29

8.1 Archive Risk Management 29

8.1.1 Risk Management in Safety-Critical Engineering 29

8.1.2 Financial Risk Management 30

8.1.3 Archive Risk Management 30

8.1.4 Workflow/Business Process Risk Modelling 31

8.1.5 User Requirements for Risk Management Tools 32

8.2 Tools for Risk Management 33

8.2.1 *Risk Framework Components* ... 33
8.2.2 *Risk Management Process* ... 35
8.3 Preservation Risk Ontology ... 37
8.3.1 *Modelling Approach* ... 37
8.3.2 *Model Definition* ... 38
8.3.3 *Classification Rules* ... 39
8.4 Risk Simulation Model ... 40
8.4.1 *Risk Generation Model* ... 41
8.4.2 *Procedure Dealing with Risk* ... 43
8.4.3 *Classification of threats/risks in digital preservation* ... 46
8.5 Preservation Metadata Model ... 47
8.5.1 *Scope* ... 47
8.5.2 *Model Definition* ... 47
8.5.3 *Relation to BPMN* ... 50
8.5.4 *Relation to Other Preservation Data Models* ... 50
8.5.5 *Specific Activities and Tools* ... 51
8.6 Preservation Metadata Service ... 52
8.6.1 *Overview* ... 52
8.6.2 *Interface to Data Providers* ... 53
8.6.3 *Interface to Consumers* ... 54
8.7 Rule-Based Decision Engine ... 54
8.7.1 *The intention behind the DAVID Rule-Based Decision Engine RTD* ... 54
8.7.2 *Decision Modelling - From Isolated Decisions to Decision Engines* ... 55
8.7.3 *Business Process Improvement through Decision Automation* ... 55
8.7.4 *Requirements for a Generalised Rule-Based Decision Engine* ... 55
8.7.5 *Implementation Variants for Rule-Based Decision Engines* ... 56
8.7.6 *Practical Use Cases for Declarative Rule-Based Engines* ... 58
8.7.7 *OMG Decision Model and Notation (DMN) - the New Standard* ... 59
8.7.8 *Conclusion and Outlook* ... 59

9 Analysis of ORF Risk Management ... 61
9.1 ORF MXF Repair Workflow ... 61
9.2 Simulation Results ... 62
9.3 Evaluation of Simulation Results by ORF ... 65

10 Conclusions ... 67

11 References ... 69

12 Glossary ... 75

13 Annex A: BPMN 2.0 Overview ... 78
13.1 BPMN Process ... 78
13.2 BPMN Collaboration ... 78

14 Annex B: ORF MXF Repair specification details ... 79

15 Annex C: Controlled vocabularies for activities, tools and risk ... 80

15.1 Specific activities ... 80

15.2 Specific tools ... 83

15.3 Risk vocabulary ... 84

16 Annex D: Preservation Metadata Service API ... 90

16.1 Definitions ... 90

16.2 Metadata Representation ... 90

16.3 Web service functionality ... 90

2 List of Figures

Figure 1: Corruption of artefact-free video during MPEG-2 to DVD conversion. 15

Figure 2: Risk framework high level component view. 34

Figure 3: ISO 31000 Risk Management Methodology. 36

Figure 4: Risk preservation layered ontology. 38

Figure 5: Preservation risk ontology – core layer. 38

Figure 6: Preservation risk ontology – sub-classing example across the domain (purple) and system (green) layers. 39

Figure 7: Risk control flow chart (omitted details of the Ad-Hoc Procedure, see Figure 8). For simplicity negative consequences are not shown, but do occur along with ‘SPOT impact’. Similarly time spend on dealing with risk is omitted, just referring to financial cost. 44

Figure 8: Ad-Hoc Control Procedure flow chart. For simplicity negative consequences are not shown, but do occur along with ‘SPOT impact’. Similarly time spend on dealing with risk is omitted, just referring to financial cost. 45

Figure 9: Entities of the preservation data model, their relations and the most important core properties. Blue entities are related BPMN entities. 48

Figure 10: SKOS approach for describing a controlled vocabulary. 51

Figure 11: Overlaying pattern. 52

Figure 12: Preservation Metadata Service and its Interfaces. 53

Figure 13: Example BPMN process within a QC process with a decision point (exclusive or gateway). 57

Figure 14: Example GUI front-end for a rule-based engine to evaluate user acceptance and user preferences. 58

Figure 15: Code snippet of a declarative definition of scheduling rules. 59

Figure 16: ORF MXF Repair workflow. 61

Figure 17: Risk Occurrence and Number of Affected Items. 62

Figure 18: Negative Consequences with and without Ad-Hoc Control. 63

Figure 19: SPOT properties affected by risks. 64

Figure 20: Ad-Hoc Control procedure usage and its overall cost. 64

3 List of Tables

Table 1: Example classification of digital damage. 15
Table 2: Key functional requirements for a risk management framework. 33
Table 3: Description of the preservation data model entities. 48
Table 4: Risk ranks in descending order. 65
Table 5: Risk ranks compared to actual (ORF) Rank. 65

4 Executive Summary

This deliverable reports on the work conducted in the DAVID project on managing the long-term quality assurance of digital Audio-Visual (AV) content. Digital preservation aims to ensure that cultural heritage is accessible for the long term. From the 20th century onwards, AV content has provided a significant record of cultural heritage, and 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 here 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. While DAVID deliverables D2.1 [Chenot 2013] and D2.2 [Hall-May 2014] has identified and analysed causes of loss, the focus here is on strategies that can be used to minimise the risk of loss. In particular, we focus on dealing with issues resulting from system errors, rather than random failure or corruption, considering the risks to the AV content as it is being manipulated by various activities in a workflow process.

Archival processes dealing with digital AV content 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 audio-visual 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, which are described in this report.

In the burgeoning domain of business process modelling (and execution), it has recently been noted that risk management can be combined with workflow specification. A conceptual risk management framework has been developed in DAVID to support a best practice approach to risk management of digital AV processes (and thus the content itself). This framework has been designed in response to the needs and requirements identified in D2.1 [Chenot 2013] and D2.2 [Hall-May 2014] as well as user input from preservation experts within the project (INA and ORF) and external participants at the first DAVID test workshop held in Vienna 2014. Most of the tools reported in this deliverable address the planning stage, where high level workflows are designed. However, it is put in context of a continual improvement cycle, based on the Deming (PDCA) cycle, involving execution of workflows, monitoring and adaptation.

The risk framework utilises a novel semantic risk model developed in the project that encapsulates domain knowledge generated in the project on known risks (and controls) associated with activities in a controlled vocabulary (also developed in the project). Not only does this provide preservation experts with a starting point for doing risk analysis, but semantic reasoning is exploited to enable suggestions on risks for the known activities. This is a part of ensuring that the risk framework is an effective help-tool to preservation experts who may not be familiar with risk specification.

The risk model developed in the project identifies risks to steps (tasks) of a workflow (as mentioned above) and classifies their impact according to a threat model (SPOT) that focuses on the essential properties of the AV asset (including, for example, identity, persistence and renderability). Controls that can be used to mitigate the risk, and the approach uses the concepts of expected loss and (conditional) value at risk. This information, along with information such as time and cost expected for dealing with risk, form input to a simulation model developed in the project. The simulation capabilities are key to the framework in order to help organisations improve cost benefit by a) identifying and understanding key vulnerabilities and b) targeting investments to address those vulnerabilities before actually executing the workflows.

Running simulations of a workflow execution allows preservation experts to determine, *inter alia*, which workflow tasks are the most affected by risk, how many times each risk occur and the time and cost spent on dealing with risk. Thus, by comparing different scenarios, such as with and without risk controls (which may come at a cost), the Return On Investment (ROI) can be analysed. For example, a preservation expert may consider the need for a new Quality Control (QC) tool that has a certain cost, and the simulation results can help determine if the ROI is likely to be positive. That is, the cost of purchasing the QC tool may be significantly less than the cost of dealing with risks that would occur without the tool. An example of risk analysis on a real-life workflow, the ORF MXF Repair workflow, is given in this report with a discussion of results obtained with the simulation model.

A preservation metadata model has been developed in the project, which allows to capture the historic creation and processing of AV assets in a common structured format. This is useful for manual analysis and audit, which is an important part of the risk management process. When workflows are designed initially, the analyses are typically based on expert judgement and limited historical data on failure. Failures are usually fixed and forgotten during times of crisis, and are often not rigorously documented. Based on execution logs generated by an execution workflow system by one of the DAVID partners, Cube-Tec, preservation metadata is extracted and made available to the risk framework. This is part of the continual improvement cycle, both in terms of adjusting the workflows themselves and to adapt the information used for initial simulations (to increase the accuracy).

Another piece of work reported on here related to the execution of workflows, is on using rule-based decision engines to automate (and optimise) the decision making in business processes. Tool support with business process modelling and integrated rule-based decision engines has the potential to help improving the robustness of this process as well as provide machine readable documentation (automated process logging) that can be used in predictive analytic models to update risk analysis estimates for future workflow executions. Several approaches to implementing rule-based decision engines, with different client-facing interfaces is considered in this report, discussing their respective pros and cons related to practical use cases.

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 work package 2) 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 two years 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*. D3.5 will report on the work in T3.3, which is due at the end of the project at M30.

This report is an update on D3.1 [Hall-May 2013]. For completeness, the content that is still relevant from D3.1 is included here. The problems and IT-based strategies for dealing with digital AV loss were well established in D3.1, so there are only minor updates to Section 6 and 7. The majority of the updates are in Section 8, regarding the conceptual risk management framework. Section 9 is new, analysing a real-life workflow with respect to risk, discussing and validating simulation results. In summary, the main updates/differences are:

- The Digital Migration (DiMi) workflow discussed in D3.1 is not included here. The focus of the work in year two has been the ORF MXF Repair workflow.
- While D3.1 focused on the risk simulation tool, we report here on a complete conceptual framework supporting user requirements identified within the consortium and the DAVID Test Workshop held in Vienna May 2014 (reported in D5.2 [Bauer 2014]) – Section 8.2.
- Updates to risk and simulation models – Sections 8.3 and 8.4.
- Preservation metadata service specification – Section 8.6.
- Additional section on rule-based decision engines – Section 8.7.
- Based on the risk specification on the ORF MXF Repair workflow, we report here on simulation results and discuss the use of risk simulation in practice (from ORF's perspective as the end-user in this case) – Section 9. Note that D3.1 only covered risk specification, not actual simulation results.

5.3 Status of this Document

This is the final version of the document.

5.4 Related Documents

Interested readers should be aware of the following related documents:

- D2.1: *Data Damage and its Consequence on Usability* [Chenot 2013]
- D2.2: *Analysis of Loss Modes in Preservation Systems* [Hall-May 2014]
- D3.1: *Initial Strategies and Risk Framework* [Hall-May 2013]

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, transferring 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 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 is 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 is 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 EC FP7 PrestoPRIME project [PrestoPRIME], which 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 WP2 of the DAVID project revealed that they rarely occur [Chenot 2013]. 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. DAVID deliverable D2.2 [Hall-May 2014] identifies three loss modes:

1. Problematic encoder: e.g., faulty encoder, inadequate encoder or the encoder is recording additional data.
2. Physical damage to files: e.g., damaged carrier, 'bit rot' and block read errors.
3. Inadequate encoder-decoder pair: e.g., ambiguous/inconsistent/mixed aspect ratio or ambiguous/inconsistent colorimetric spaces.

Problems that are a result of system error are potentially much more damaging, opposed to random failure or corruption (such as 'bit rot'), as they 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, which is addressed in D2.2 [Hall-May 2014] in more depth. Consider, for example, the following examples of problems that are a result of system error:

- 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.

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.

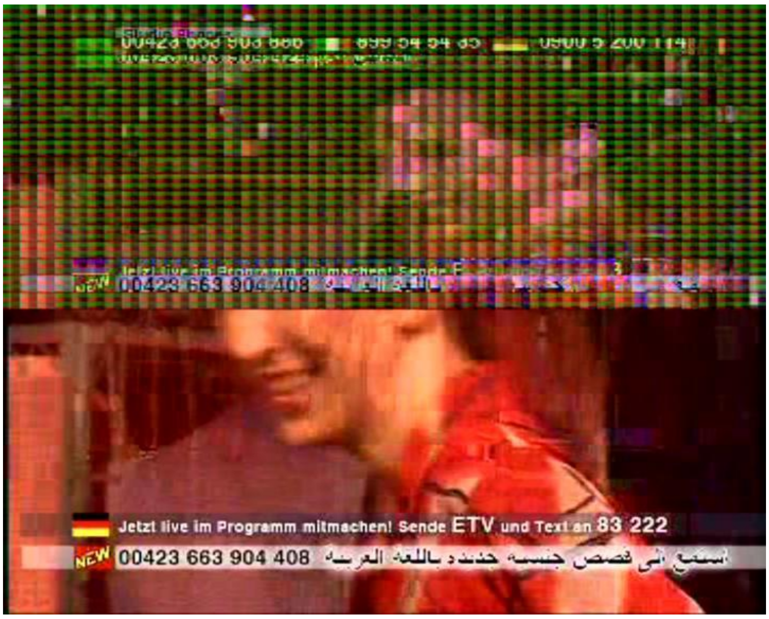

NEW
Jetzt live im Programm mitmachen! Sende ETV und Text an 83 222
00423 663 904 408

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 DAVID WP2 analyses [Chenot 2013, Hall-May 2014], 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

Process	Problem	Cause
		Artefacts existed in source prior to ingest
		Playout malfunction
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 [Chenot 2013]. 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.

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.

7 IT-based Strategies Against 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. Section 7.1 discusses how risk to digital AV content can be addressed, and Sections 7.2 and 7.3 discusses the IT-based preservation strategies taken by INA and ORF, respectively.

7.1 Avoiding, Mitigating and Recovering from Digital AV Loss

There are essentially three approaches for addressing (treating) risks related to the loss of digital AV content:

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 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 (HDDs), Solid-State Drives (SSDs), 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 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.

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.

The 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).

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, MPEG-4 AVCI, 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.

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 is 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.

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.

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] and MXF Legalizer [MXF Legalizer].

The Planets 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.

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, LTO6), or on disk (HDD).

Security of data on the frequently-accessed 3466 tapes stored into the automated tape library (4.1 PetaBytes, 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 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 them 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 digital 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 used to be stored on a number of RAID 6 appliances in a remote location aimed at restarting the activity in case of a disaster. Location of Recovery copies has changed in 2014, and two data tape library (LTO-6) are progressively replacing the energy-demanding RAID 6 appliances. 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 (e.g. LTO-3). Although not strictly speaking a file, INA also keeps 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 in less than 48 hours the core capacity for delivering contents to customers. 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. INA are progressively modifying the Disaster Recovery approach into a dual-location load-balancing mode, where requests will be served from either location (Main, Recovery), depending on availability and load.

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, rather than modifying the files, it is often easier to adjust the tools and procedures to allow using the affected files anyway.

7.2.5 Comparison with Analogue Workflows

The situation in the digital workflows 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. 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. Even after digitisation, destruction of the analogue carriers is strictly limited to cases where the original is damaged beyond possible repair (e.g. vinegar syndrome), or when it is proven that several better copies are already available (e.g. VHS copies).

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 tapes 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

The main purpose of the ORF-Archive has a big difference to the INA-Archive; it is mainly a "production-support archive", so constant availability and fast access (within minutes for News-Productions) are the key elements. Therefore, ORF does not have "special" archive-formats, but use the actual production formats (MXF D10 and MXF XDCAM HD422) instead to avoid delays in access due to file-conversion.

7.3.1 Physical Security of AV Data

ORF started in mid-2012 with pure file-based archiving; since then all production-units at ORF has been switched over to a "file-only"-digital workflow. All production-units preserve their content in a file-based storage system ("ESYS", which is short for Essence-System). ESYS is triggered and managed by the 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®).

7.3.2 Backups of AV Data

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 the internal 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; but 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 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 ORFs Digital Migration Project (DiMi) ORF had to decide to no longer stick to the old rule of "keeping the source", since due to commercial reasons the space is be needed otherwise. So after the final checks in the DiMi workflow all analogue (and digital) sources will be disposed of.

7.3.3 Security of AV Data and File Conformance

In addition to the Quality Check-Routines in ESYS (automated; general check of incoming files; conformance of files is checked against profiles to prevent inconsistent file structures in the ORF 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:

Class 1: The source of the files is known, no detailed testing is performed.

Class 2: 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.

Class 3: Unknown file sources: an automatic, or if not possible, also manually executed transcoding tries to make the files useable.

This approach minimises 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 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.

7.3.4 Comparison with Analogue Workflows

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), ORF expects a similar or higher rate of “Preservation-security” for contents in 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 our expectation) more reliable, as they are mostly automated, while the earlier ones where always carried out manually.

In summary the overall expectation here is that the 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 (Quality of 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. 300,000 hours of AV-content from IMX and Digital Betacam to MXF D10 OP1a; those will be stored then in the ESYS-system.

8 Conceptual Risk Management Framework

As discussed in Section 5.2, this section contains the most significant updates since D3.1 [Hall-May 2013]. Section 8.1, discusses archive risk management, ending with a set of user requirements that underpin the tools for risk management discussed in Section 8.2. This latter section offers an overview of the tools and their relationships, which are further discussed in respective sections that follow (8.3 - 8.7).

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. In general terms, we can say that the key objective is to ensure long-term preservation of digital AV content, i.e., avoid damage and ensure that it can be accessed in the future.

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.

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 0. 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.

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 8.4. 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.

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.

8.1.5 User Requirements for Risk Management Tools

In archive management, the key actor we are addressing in this report is the preservation expert / specialist, who is responsible for designing workflows for managing and processing digital AV content. While we have discussed the motivations for performing risk management already, we can summarise here some key purposes of a risk management framework in the context of digital preservation:

1. Helping preservation experts develop new workflows, especially the early stages of development. Note that the purpose of the framework is not to replace the preservation experts, but to be a value-added tool to assist them.
2. Helping preservation experts optimise workflows (in terms of cost effectiveness and security), considering also trade-offs where too many corners are cut, which leads to high risks. In terms of risk mitigation, we here refer to reducing the likelihood or impact of risks.
3. Helping preservation experts communicate and justify decisions about choices for elements in workflows. This may be related to arguing expected financial Return On Investment (ROI) of putting in place certain risk mitigations, for example.
4. Helping organisations change their processes, as the risk of change is typically seen as very high, which inhibits change. That is, improving workflows while ensuring that they are secure. Risk simulations in this context is of particular value, to be able to simulate different “what if” scenarios to determine the likely outcomes of potential changes.

From an organisational point of view, the reasons to perform risk management can be summarised as:

1. Preservation workflows can be large and complex, so there can be too many variables and options for preservation experts to consider simultaneously. As mentioned above, it is again emphasised that the risk framework is a support tool, not a replacement.
2. Risk information is typically in experts’ heads, which is itself a risk from the organisation’s point of view. The risk framework ensures that the knowledge is captured and retained, and is readily available should the organisation be subject to an audit.
3. Improve cost-benefit by a) identifying and understanding key vulnerabilities and b) targeting investments to address those vulnerabilities.
4. Move away from “firefighting”. That is, organisations may spend more time dealing with issues rather than preventing them in the first place. Risk management is key to prevention; *i.e.*, spending more time in the planning stages to save time and cost on dealing with issues in the future that could have been avoided.

It is important to note that the key user of the risk management framework in this context is not a risk expert. They are preservation experts, but they will be acutely aware of a wide range of potential issues concerning the preservation workflows they manage. However, explicitly managing risk may be very unfamiliar and it is important that the risk management framework is suitably designed to aid preservation experts (rather than simply being a risk registry).

At a high level, we include here some key functional requirements of a risk management framework for digital AV preservation (below in Table 2). These requirements have been captured from within the DAVID project (i.e., the project partners, especially ORF and INA), as well as from participants at the DAVID Test Workshop in Vienna in May 2014 [Bauer 2014].

Table 2: Key functional requirements for a risk management framework.

Category	ID	Description
Risk specification	R01	Embedded with familiar workflow specification tools, to specify risks for individual tasks in a workflow.
	R02	Maintain a register/repository of all known risks, and be able to export this register/repository for reporting purposes.
	R03	Get assistance for specifying risks for known tasks.
Simulation	R04	Specify parameters to enable simulation of running workflows, such as the duration of tasks, frequency of risk occurrence and cost of dealing with the risks.
	R05	Be able to simulate different scenarios, based on, for example, using risk controls (to mitigate risk) or not, different workloads (e.g., throughput of files).
	R06	Be able to visualise and compare results from different simulation scenarios, and export the results for reporting purposes (similar to R03).
Monitoring and adaptation	R07	View monitoring information from workflow processes that have executed.
	R08	View meta-data about workflow process executions, to analyse risk.
	R09	Adapt workflows or simulation parameters according to monitoring information.

8.2 Tools for Risk Management

Based on the above requirements, we can identify 4 key elements of the risk framework:

Workflow Designer: encapsulating all the functionality required to create and edit workflows.

Risk Editor: to specify risks, representing both the general risk repository and specific risk instances for specific workflows. Note that risk specification could be embedded in the workflow designer too.

Simulation Centre: to run simulations of workflows annotated with risk information. Require risk details such as probability of occurrence, not just that there is a risk.

Risk Feedback Centre: to be able to view and analyse workflow execution data (and preservation meta-data), to further analyse risk and adapt simulation models.

The various components of the framework are further discussed in Section 8.2.1. Thereafter, Section 8.2.2 discusses how these components fit into a best practice process for risk management. In generic terms, we refer to this risk framework as a Business Process Risk Management Framework, or BPRisk as a short term, which is used for convenience throughout the remainder of this section.

8.2.1 Risk Framework Components

A high level architectural component diagram of the conceptual risk management framework is provided in Figure 2, below, addressing the requirements discussed above in Section 8.1.5. This diagram gives an architectural overview, showing the relationship between tools discussed in this section. Note that the components in blue are tools developed in the DAVID project and the components in purple are previously existing tools that are either directly a part of the framework (embedded), or part of the preservation risk management process (3^{rd} party tools or services). Brief descriptions of each component is given below, clarifying their role and references to respective sections further in this report is given for more information.

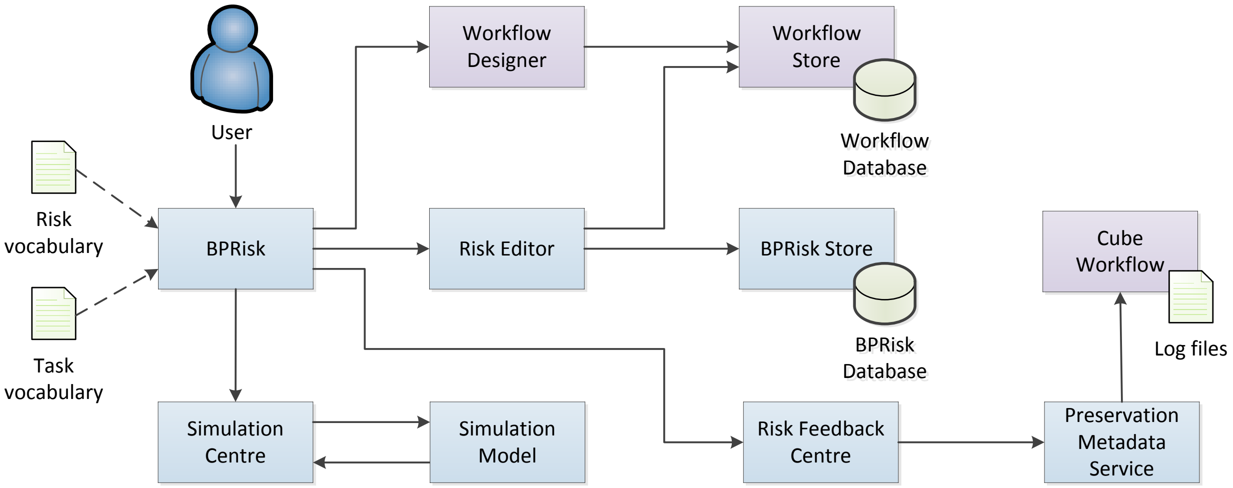


Figure 2: Risk framework high level component view.

BPRisk: The Business Process Risk Management Framework (BPRisk), the main entry point for the user. This is depicted as a web application from which the user can access the functionalities of the framework, e.g., to launch the workflow designer, open existing workflows, specify risks, run and view risk simulation results, etc. Figure 2 also shows two vocabularies used, one for known domain-specific risk and one for domain specific tasks, which are detailed in Annex C (Section 15 on page 80).

Workflow Designer: The workflow designer should offer CRUD functionality for (preservation) workflows. There are several existing, mature, tools for this, supporting the well-known BPMN 2.0 standard [BPMN], such as Signavio [Signavio] and the jBPM Designer [jBPM]. In terms of embedding a designer and extending it for risk specification, jBPM is a free, open source, software that enables this. For readers who are unfamiliar with BPMN, please see Annex A (Section 13 on page 78) for a brief overview.

Workflow Store: This is a component to persist any workflows created, updated or imported. Existing tools, such as jBPM come with multiple persistence options and a RESTful API for accessing and managing the workflows.

Risk Editor: As described above, this component is responsible for allowing users to specify risks. We particularly refer to requirements R02 and R03 in Table 2 on page 33, and Section 8.3 describes a preservation risk ontology for capturing risk information, as well as enabling a mechanism for suggesting risks to users for known tasks. The latter point is a part of an aim in DAVID to make the domain knowledge generated within the project available to the wider community. This also relates to the efforts on preservation metadata, which is discussed in Section 8.5.

BPRisk Store: This is a component for persisting risk specifications and risk simulation results (a connection from the Simulation Centre has not been depicted in Figure 2 for the sake of simplifying the diagram).

Simulation Centre: This is a component for managing the running of simulation models for workflows annotated with risk information, and does any required pre and post processing of data (input to the models and output from the models).

Simulation Model: A risk simulation model that the Simulation Centre can execute, as described above. Section 8.4 goes into further detail about the risk simulation model developed in the DAVID project, and results from using this model are discussed in Section 9.

Risk Feedback Centre: A component in the Risk Framework for getting data from real workflow executions that can be used to a) analyse the workflow execution data and b) to modify/adapt/calibrate the risk details and simulation to improve the accuracy. For the former, Sections 8.5 and 8.6 go into detail of the preservation metadata model and service developed in the DAVID project.

Preservation Metadata Service: We have referred to this component already above, which is a service that implements the preservation metadata model described in Section 8.5. This service interacts with an external component, Cube Workflow, a tool from Cube-Tec to execute workflow processes. Based on logged information about the workflow executions, preservation metadata is extracted, which the Risk Feedback Centre consumes (as described above). Further details about the preservation metadata service are given in Section 8.6.

Cube Workflow: An external software system by Cube-Tec International (one of the DAVID partners) for executing workflows. This software is based on jBPM, using two artefacts in particular, the jBPM Designer and the jBPM Workflow Engine. This is a framework that also includes the rule-based decision engine work discussed in Section 8.7, which aims at automating (and optimising) the decision making in business process executions.

8.2.2 Risk Management Process

The risk framework should support a process that is based on best practices and is aligned with risk management methodologies and the needs of preservation experts. For the latter we already have discussed some functional requirements in Section 8.1.5 on page 32, but not really going into actual usage patterns. There is a focus on the planning aspects regarding risk management, which has been discussed above in Sections 7 and 8.1, but we do need to consider the wider context as well.

There are several risk standards and methodologies, but it is not within the scope here to discuss them in detail. However, we will make reference to one in particular here, ISO 31000 [ISO31000, 2009], to show how it aligns with a best practice approach proposed here based on the Deming cycle. The Deming cycle is a four-step iterative method commonly used for control and continuous improvement of processes and products, and is key to, for example, ITIL Continual Service Improvement [Lloyd 2011]. In general terms, risk management is a part of continual improvement of processes – preservation workflows in this context.

As mentioned above, the Deming cycle consists of four iterative steps: Plan, Do, Check and Act. For this reason it is also commonly referred to as the PDCA cycle. The ISO 31000 [ISO31000, 2009] risk management methodology is depicted in Figure 3, below, which depicts the various stages from ‘establishing the context’ to ‘treatment’ (of risk) that is also cyclic.

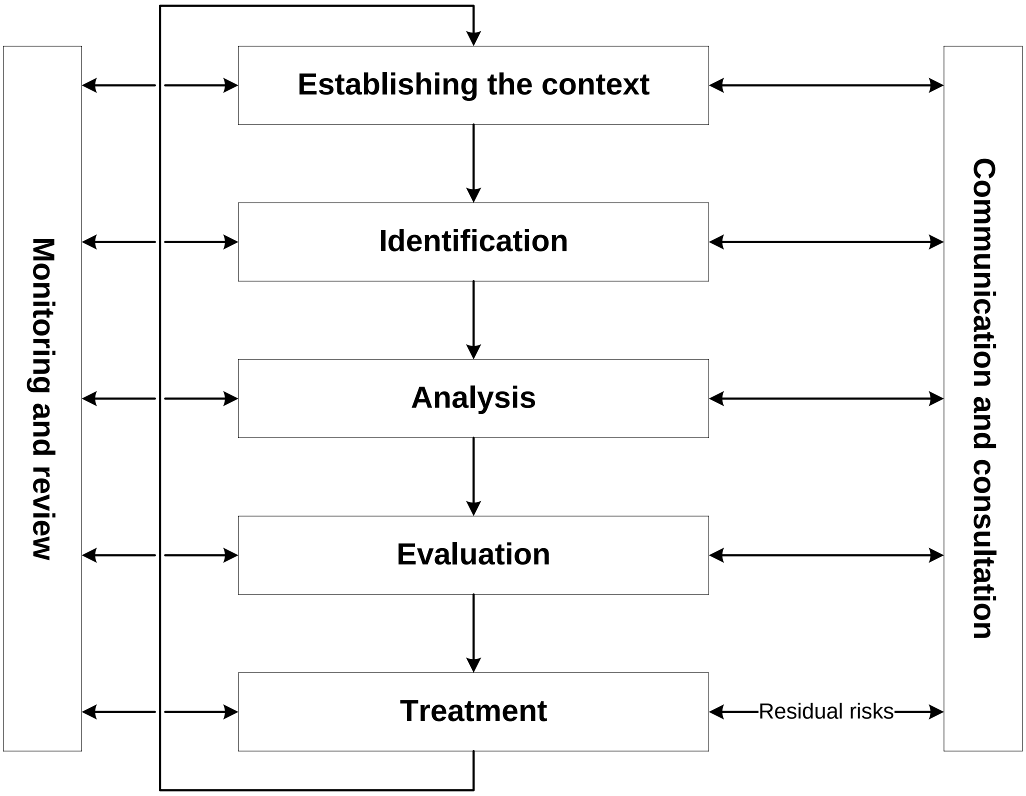


Figure 3: ISO 31000 Risk Management Methodology.

Given the risk framework components described above in Section 8.2.1, each of the four stages of the Deming cycle is covered below from the perspective of what a user (preservation expert) would do.

Plan (covers the 'establishing the context' and 'identification' stages of ISO 31000):

- The user designs a high level business process workflow using the Workflow Designer. The workflow is persisted in the Workflow Store.
- The user identifies and adds information about risks and controls to tasks in the workflow using the Risk Editor (also updating the risk repository). The risk information is persisted in the BPRisk Store.
- The user uses the Simulation Centre to select a workflow for simulation and provides additional simulation parameters (see Section 8.4 for further details).
- The user uses the Simulation Centre to run simulations and views the results.
    - The user may compare the results with a previous run or another workflow (which may have been an earlier revision for the same purpose).
    - Results are persisted in the BPRisk Store.
- The user may now proceed with further offline analysis (**Act**) or with online execution (**Do**)
    - Further offline analysis may be: a) different risk specifications for a workflow and b) different simulation scenarios (rate of files, control options, risk probabilities, etc.).

Do (covers the 'analysis' stage of ISO 31000):

- A Cube-Tec employee builds a concrete (technical) business process using the Cube Workflow designer, based on the high level workflow[1].
    - imports and updates the BPMN, substituting abstract activities with 'real' tools where required.

[1] Note that Cube-Tec does work with clients in this way, creating concrete business processes from the more high-level workflows specified either by or in collaboration with their clients.

- User executes the concrete business process in Cube Workflow.
- User monitors the business process execution (**Check**), integrated on a Dashboard in the Risk Feedback Centre via the Preservation Metadata Service.

Check (covers the 'evaluation' stage of ISO 31000):

- User retrieves workflow process execution data via the Preservation Metadata Service in the Risk Feedback Centre.
    - Data retrieved from the Preservation Metadata Service is gathered and processed from Cube Workflow execution logs.
- User uses the Risk Feedback Centre to visualise the data in order to analyse risk occurrences.
- User may decide that some actions are needed (**Act**).

Act (covers the 'treatment' stage of ISO 31000 as well as feedback to the previous stages):

- User takes one of several possible actions based on the feedback from the real process execution and/or simulation results analysis:
    - Redesigns the workflow, e.g., adding, removing or reordering activities.
    - Changes risk assumptions, adding, removing or updating risks and their parameters.
    - Changes risk treatment, adding, removing or updating controls and their parameters.
    - Changes simulation parameters / runs new simulations.
- User re-runs simulation with the new input (**Plan**) or enacts the offline changes in the real business process and continues execution (**Do**) and monitoring (**Check**).

8.3 Preservation Risk Ontology

As discussed previously in Section 8.1.5 on page 32, a preservation risk ontology has been created to a) allow users to maintain a repository of risks and to b) receive assistance for risks relating to known tasks.

8.3.1 Modelling Approach

The risk preservation ontology represents information related to risks, controls and activities (BPMN tasks). This representation allows flexibility and extensibility of the risk model. It can be easily published (e.g., as a set of OWL files), can be extended in unexpected ways, can be combined with other ontologies, and it allows the use of off-the-shelf tools such as triple stores and semantic reasoners.

The approach to building the ontology is based on work done in SERSCIS project [Surridge 2012]. The authors use a layered, class-based ontology model to represent knowledge about security threats, assets and controls. Each layer inherits from the layer above. The CORE layer describes the relationships between a central triad (threat, asset, control). A domain security expert subclasses each of these core concepts to create a DOMAIN[2] layer. A system expert further subclasses the generic concepts to specialise them for the system of interest, creating the SYSTEM layer.

The same, layered, ontological approach has been taken in DAVID, but the core ontology is slightly different. While, in SERSCIS, the triad in the CORE layer includes Asset, there is only one asset of value in this context – the digital AV object, whose value can be degraded by the effects of certain processes applied to it during the preservation lifecycle (i.e., workflow tasks such as ingest, storage and transcoding). The term Threat used in SERSCIS can be understood as Risk in this context. Therefore, the CORE layer in DAVID comprises a triad of Risk, Activity and Control.

The three layers of the preservation risk ontology can be seen below in Figure 4, with some examples of activities, risks and controls.

[2] In SERSCIS, the DOMAIN layer is called the GENERIC layer. The term is changed here as it is more intuitive in this application context.

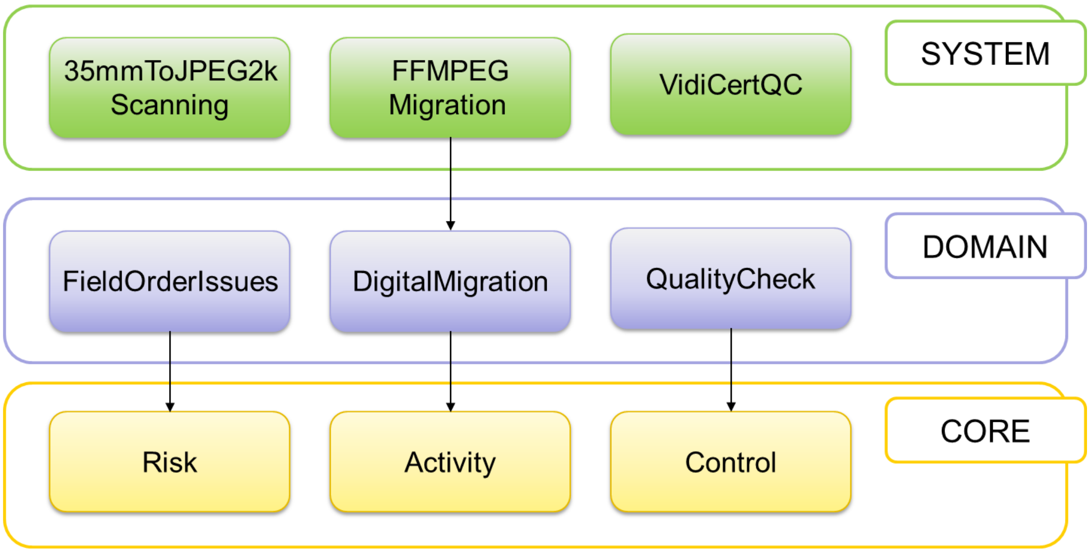


Figure 4: Risk preservation layered ontology.

8.3.2 *Model Definition*

In DAVID, the model focuses on the Activities in the preservation lifecycle and the Risks that are inherent in their execution. Controls can be put in place to block or mitigate these Risks. The CORE layer comprises risk, activity, control as seen above, as well as basic relationships such as 'Risk threatens Activity' and 'Control protects Activity'. However, the relationship between Control and Risk is established via SPIN rules (see the following section), to determine the appropriate relationship. That is, a Risk is only considered Mitigated if an appropriate Control is in place. This is illustrated below in Figure 5.

The DOMAIN layer has been developed in collaboration between IT Innovation, ORF and JRS in the DAVID project, and describes the general preservation activities, risks and controls that are known. These are modelled as sub-classes, which is illustrated below in Figure 6. The DOMAIN level entities are coloured purple, while an example of a SYSTEM level entity (FFmpeg Migration) is depicted in green. The SYSTEM layer would be populated by the users of the risk management framework – when they build a workflow of specific Activities and associate Risk to them.

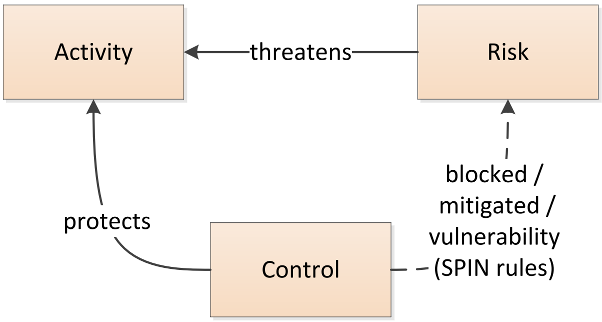


Figure 5: Preservation risk ontology – core layer.

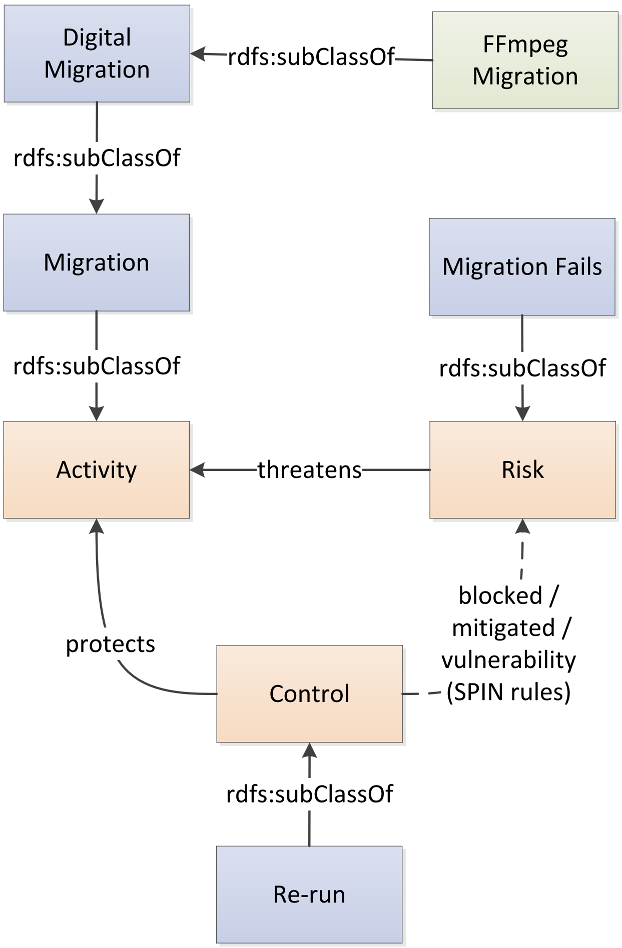


Figure 6: Preservation risk ontology – sub-classing example across the domain (purple) and system (green) layers.

8.3.3 Classification Rules

The relationship between risks, controls and activities are encoded as RiskClassification rules using SPIN. Running inferencing over the model automatically does the classification. For example, the following SPIN rule classifies an instance of the risk FieldOrderIssues, which threatens the activity Transcoding, as *blocked*, if the control ChangeTranscodingTool is present:

```
CONSTRUCT {
   ?r a dom:BlockedRisk .
}
WHERE {
   ?a a act:Transcoding .
   ?r a dom:FieldOrderIssues .
   ?c a dom:ChangeTranscodingTool .
   ?r core:threatens ?a .
   ?c core:protects ?a .
}
```

The SYSTEM layer will be developed so that it subclasses the DOMAIN layer for a specific organisation using the risk framework, as seen above in Figure 6. This should specify the kind of activity in the preservation workflow of interest, e.g., subclass Scanning as 35mmToJPEG2kScanning. The workflow-specific risks can be automatically generated using SPIN. For example, the following is a generic SPIN rule to generate all risks:

```
CONSTRUCT {
  ?uri a owl:Class .
  ?uri rdfs:subClassOf ?gr .
  ?uri rdfs:subClassOf _:b0 .
  _:b0 a owl:Restriction .
  _:b0 owl:onProperty core:threatens .
  _:b0 owl:someValuesFrom ?sa .
}
WHERE {
  ?sa (rdfs:subClassOf)+ act:Activity .
  ?sa rdfs:subClassOf ?ga .
  ?gr rdfs:subClassOf core:Risk .
  ?gr rdfs:subClassOf ?restriction1 .
  ?restriction1 owl:onProperty core:threatens .
  ?restriction1 owl:someValuesFrom ?ga .
  FILTER NOT EXISTS {
    ?uri rdfs:subClassOf _:0 .
  } .
  FILTER STRSTARTS(str(?sa), "http://www.david-preservation.eu/bprisk#") .
  BIND (fn:concat(STRAFTER(str(?gr), "#"), "_", STRAFTER(str(?sa), "#")) AS ?newclass) .
  BIND (URI(fn:concat(fn:concat(STRBEFORE(str(?sa), "#"), "#"), ?newclass)) AS ?uri) .
}
```

This rule finds all activities in the SYSTEM layer and creates a workflow-specific risk for each of the generic risks that threaten the activities' parent class. The name of the workflow-specific risk is generated by concatenation of the generic risk name and the workflow-specific activity name. For example, if FFMPEGMigration is a subclass of the activity DigitalMigration, then the risk generation rule would create the following workflow-specific risks based on the domain-level risks that threaten DigitalMigration:

```
http://www.david-preservation.eu/bprisk#IncorrectInformationMetadataAddition_FFMPEGMigration
http://www.david-preservation.eu/bprisk#InformationMetadataLoss_FFMPEGMigration
http://www.david-preservation.eu/bprisk#QualityLoss_FFMPEGMigration
```

As noted above, the Activity and Risk (and control) model is based on the taxonomy generated by JRS in collaboration with IT Innovation and ORF. See the preservation metadata model below in Section 8.5 and the controlled vocabularies in Section 15 (Annex C).

8.4 Risk Simulation Model

Previously, in D3.1 [Hall-May 2013], we presented a risk management model as a static model which takes into consideration the probabilities of risk occurrence and allows us to make assessment of different risk measures for a given workflow on average. The drawback of this model is its deterministic nature, i.e., it does not allow us to investigate different scenarios and simulate risk occurrences as in a real life. To overcome this drawback, a stochastic risk management model was developed. This model allows users to simulate different scenarios and to produce confidence intervals for different risk measures if required by means of Monte Carlo simulations. Moreover, the stochastic model allows to get answers for what if scenarios and can be used both during planning and operation stages.

The proposed stochastic risk management model consists of two main parts:

- Risk generation model.

- Procedure for dealing with risks.

Below we describe each of these parts in detail.

8.4.1 Risk Generation Model

The stochastic risk generation model is based on simulating a given workflow, where risks associated with each task take place based on risk occurrence probabilities and dependencies between risks. Dependencies between risks can be within a single task or in consecutive tasks. Below we state all data about a workflow that is needed, divided into the following categories for convenience: general, workflow-related, risk-related, control-related[3] and other simulation parameters.

General data:

- The purpose of the workflow under consideration, that is what the workflow does, inputs and outputs to/from the workflow
- The objectives of risk analysis for this workflow

Workflow-related data:

- List of tasks in the workflow, a short description of each task, and how tasks are connected between each other
- Duration of each task
- Decision points in the workflow, and based on records or previous experience how often each decision are usually made at each decision point (e.g. at decision point 1, D1 will be made approximately 90% of times and D2 10% of times).

Risk-related data (<u>for each task</u>):

- List of risks (threats) which can take place and their descriptions
- Any dependencies between the risks in the same task and/or risks from different tasks, e.g. can the risks in the same task occur together?
- Frequency of each risk occurrence either from records or estimated based on the previous experience; frequencies of more than one risk taken place in a task if relevant; any changes in frequency of occurrence of some risk in the task if other risk in the same task took place
- For each risk
  - Causes
  - Probability (frequency) of occurrence
  - Detection level (if known)
  - Consequences for the workflow (previous task, next task and the whole workflow)
  - Negative impact on workflow measured in monetary values, percentages or some impact scale
  - Affected SPOT properties (explained further below)
- If combination of risks can occur:
  - Frequency of combined occurrence
  - Multiplication factor

Control-related data (<u>for each risk</u>):

- Is anything done on the fly (Ad-Hoc Control)? If yes, is the Ad-Hoc Control procedure covered by overhead? How effective is the Ad-Hoc Control procedure (a value for ‘Expected Success’ would be provided)?
- List all other control procedures dealing with this risk and their effectiveness
- If more than one procedure dealing with risk is available, describe conditions when different procedures are activated
- List costs associated dealing with risk and time spent on dealing with risk

[3] Control-related data describes mechanisms for addressing risks, e.g., to avoid the risk entirely, reduce the likelihood or reduce the impact should the risk occur.

- Describe how negative impact is reduced when Ad-hoc or/and other control procedures are applied

Other simulation parameters:

- Number of items to be processed through a workflow
- Annual throughput
- Number of items to be processed during a day, week, month, year

The Risk Occurrence probability is calculated based on the Estimated Frequencies (EFs) of risks provided by a user. At the moment, it is assumed that EFs of risks are based on a pre-defined annual throughput of a given workflow, and therefore a risk occurrence is simulated on item basis. If a pre-defined annual throughput of the workflow changes, the new estimated frequencies can be updated using Estimated Frequency Factor (EFF) provided by a user. For example, if we are interested in processing *n* items per month, which is equivalent to *12*n* items per year, then the EF for risk *i* in task *j* can be calculated as:

$$\text{EF\_new}(i,j)=12*n*\text{EFF}(i,j), \qquad (8.2.1)$$

where EFF(i,j) is Estimated Frequency Factor for risk *i* in task *j*.

EF_new(i,j) will be represent a frequency per year in this form.

The probability of risk occurrence based on EF per item can be calculated as

$$P(\text{risk for 1 item}) = \frac{EF}{nItems}, \qquad (8.2.2)$$

where EF is estimated frequency based on a pre-defined throughput for a given workflow in pD (per Day), pW (per Week), pM (per Month) and pY (per year).

nItems corresponds to a number of Items which can be processed in a day, week, month, year based on this pre-defined throughput.

If the throughput if items is changed than instead of EF, EF_new (see formula (8.2.1)) will be used together with new values pY, pM, pW and pD.

By the nature of dependency between risk occurrences, all risks can be divided in the following groups:

1) Risks do not have any known dependency between each other and it is assumed that they cannot happen at the same time.
2) The frequency of one risk in a given task changes temporally if another risk in this task took place. In this case these risks can occur at the same time. Otherwise both of the risks can happen only one by one (mutually exclusive).
3) The frequency of ***Risk A*** in the next task changes temporally if ***Risk B*** in the previous task occurs. This situation will be modelled as follows: if ***Risk B*** took place, then the frequency of ***Risk A*** is changed accordingly (for a single run of the workflow).
4) The risks can happen together. In this case the frequencies of each risk and frequency of risks occurring together are used to simulate such a situation. In this case Estimated Frequency means that only a given risk took place, Frequency of combined shows the joint frequency of risks.

The identification of risk generation group will be done automatically checking corresponding values in order of priority:

1. Multiple Risk-Entry per Task
2. Multiplication Factor
3. Frequency of combined

A Detection Level parameter allows us to simulate whether a risk was detected or not. If risk is detected, then procedure dealing with risk will be put in action (see the next section). Otherwise we mark affected SPOT properties and record level of negative consequences (NC).

An example of risk specification and related simulation configuration is given in Annex B (Section 14 on page 79). Based on the above input data, the raw outputs from risk generation model are:

- Flag indicating whether risk *i* in task *j* took place (1- for risk occurrence, 0- no risk)
- Name of risk
- And in case Flag=1, how many items were affected by this risk
- Flag indicating whether risk *i* in task *j* were detected for each risk occurrence for a given task and run

Based on the above raw outputs, many statistics can be calculated and presented. See Section 9 on page 61 for results on a real-world workflow, the ORF MXF Repair workflow.

8.4.2 Procedure Dealing with Risk

The risk simulation tool implements a procedure of dealing with risk that comprises two types of controls: ***Ad-Hoc Control*** and ***Active Control***. As we illustrate below in Figure 7, these controls only apply if a risk is actually detected. If a risk is detected, then the Ad-Hoc Control procedure is started. Figure 8 illustrates what occurs in the Ad-Hoc Control procedure, which we will return to below. In short, the Ad-Hoc Control procedure signifies the process for rectifying issues, such as repairing files that may have been damaged.

Active Control is applied only if a cost spent on an Ad-Hoc Control procedure is higher than a pre-defined value, as illustrated in Figure 7. This is generally only likely to be issued for large and significant issues and could be, for example, re-training staff or allocating more resources. This type of control would typically incur additional cost and time to put into place. However, note that Active Control is not necessarily available for all tasks/risks. In this case only the Ad-Hoc Control procedure is applied for dealing with those risks.

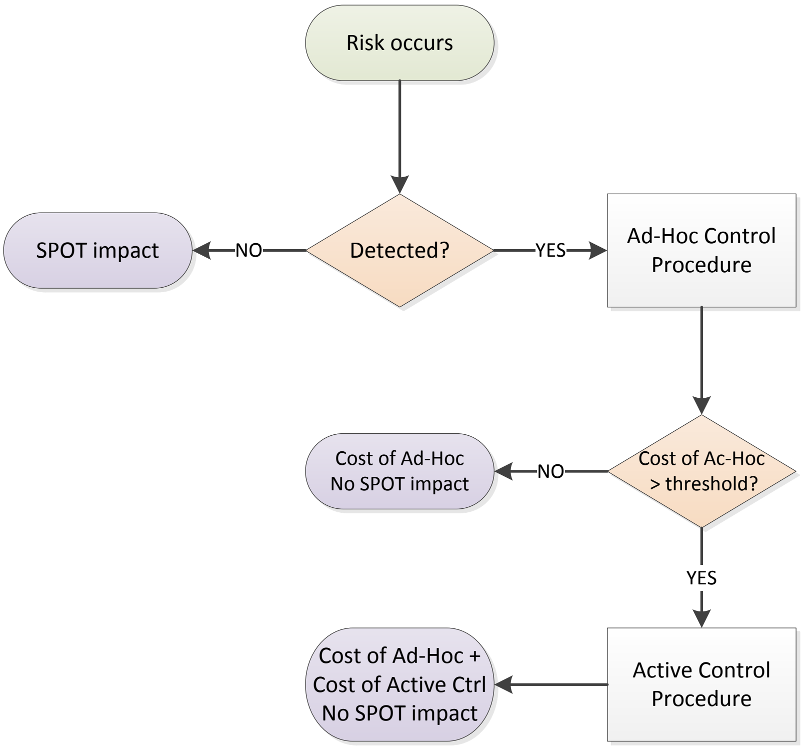

Risk occurs
Detected?
NO
SPOT impact
YES
Ad-Hoc Control Procedure
Cost of Ac-Hoc > threshold?
NO
Cost of Ad-Hoc No SPOT impact
YES
Active Control Procedure
Cost of Ad-Hoc + Cost of Active Ctrl No SPOT impact

Figure 7: Risk control flow chart (omitted details of the Ad-Hoc Procedure, see Figure 8). For simplicity negative consequences are not shown, but do occur along with 'SPOT impact'. Similarly time spend on dealing with risk is omitted, just referring to financial cost.

The details of the Ad-Hoc Control procedure are illustrated below in Figure 8. It can be covered by Overhead[4] or not. If not, it will result in cost and time spent with dealing with the risk. However, if the Ad-Hoc Control procedure is covered by Overhead and has 100% Expected Success of ad-hoc counter-measures, then a) the risk doesn't have any effects on the assets properties (SPOT model) or Negative Consequences and b) there are no (extra) costs associated with dealing with the risk.

If the Expected Success of the Ad-Hoc Control procedure is not 100%, then, based on the number of items to be processed through the workflow and Excepted Success rate, additional Ad-Hoc Control procedures are performed as follows. Note that, in this case, the Ad-Hoc Control procedure is performed until Negative Consequences is zero or near zero[5].

1. For up to 10 items: the 1st attempt of the Ad-Hoc Control procedure always successful independent of the Expected Success rate provided for the respective risk.
2. From 11 to 100 items: 2nd attempts will be always successful if Success rate is 50% or above, 3 attempts will have to be made otherwise to achieve success.
3. From 101 to 1000 items: 3rd attempt will be needed to reduce the Negative Consequences to zero.

In general, the number of attempts needed is equal to the order of the number of items to be processed via the workflow. Let's take an example:

> A risk is detected for 500 items, and we have a 90% Expected Success rate for this risk.
>
> This means the 1st attempt will resolve the issue for 450 items → 50 items remaining.
>
> The remaining 50 items are subject to a 2nd attempt → 5 items remaining.

[4] Overhead is a term used here for either a budget or a percentage of resources set aside *a priori* to cover the cost of dealing with issues.

[5] Negative Consequences assumed here is an "impact" value in the range of 1 – 5 (low to high).

The remaining 5 items are then subject to a 3rd and final attempt, and for up to 10 items, we have modelled the attempt to have a 100% success rate, regardless of the value provided for the Expected Success rate.

For each attempt, time and cost is accumulated, and it is this sum that is subject to the Active Control check; i.e., if this exceeds some pre-defined threshold.

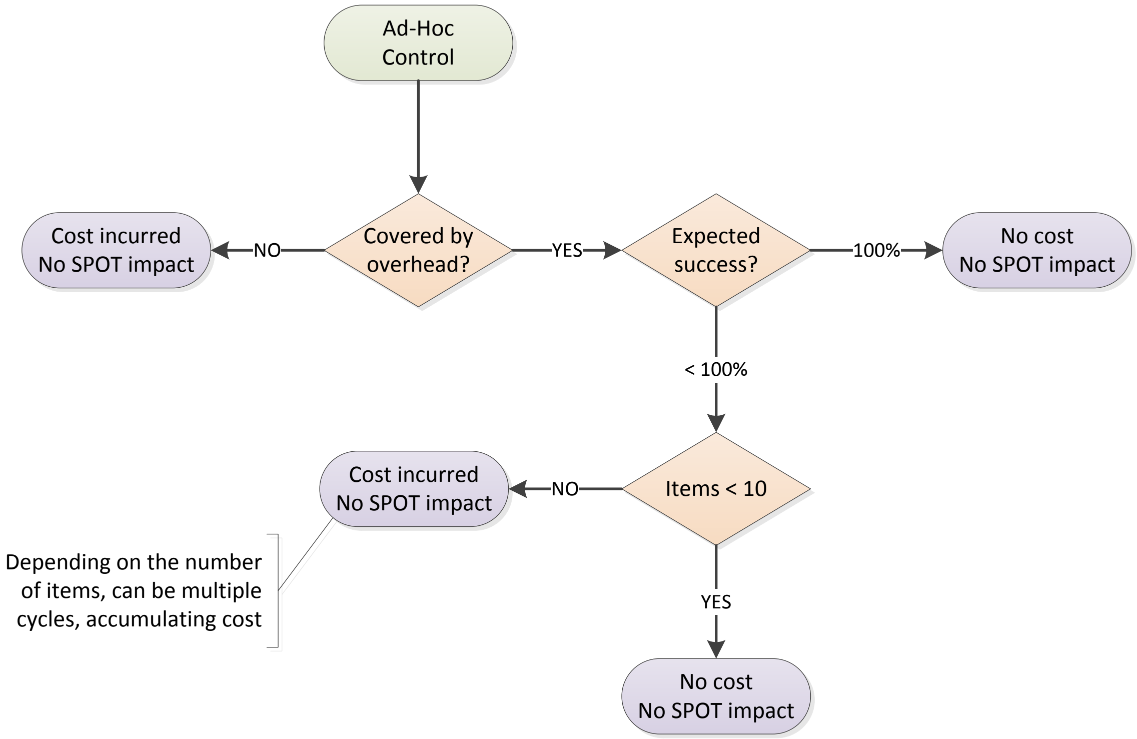

Ad-Hoc Control
Covered by overhead?
NO
Cost incurred No SPOT impact
YES
Expected success?
100%
No cost No SPOT impact
< 100%
Items < 10
NO
Cost incurred No SPOT impact
Depending on the number of items, can be multiple cycles, accumulating cost
YES
No cost No SPOT impact

Figure 8: Ad-Hoc Control Procedure flow chart. For simplicity negative consequences are not shown, but do occur along with 'SPOT impact'. Similarly time spend on dealing with risk is omitted, just referring to financial cost.

The general formula for calculating cost of the Ad-Hoc Control procedure is

$$Cost = TAR * CAR * \sum_{i=1}^{k} n_i \ , \qquad (8.2.3)$$

where TAR is time needed for dealing with risk for one affected item,

CAR is a cost associated with dealing with risk for 1 hour TAR,

k is a number of attempts of the Ad Hoc Control procedure calculated as above based on a number of items passed through the workflow,

n_i is a number of affected items after each Ad-Hoc effort calculated as a product of number of affected items before *i*th attempt and (1 - Success rate of Ad Hoc) in decimal points.

Active Control is applied only if a cost spent on Ad-Hoc Control is higher than a pre-defined value (CACS). Negative consequences (NC) will be reduced by a percentage in Expected control Success column. Active Control can be not available for all tasks/risks. If Active Control is available then a check is needed for its activation. Activation of Active Control is possible after any number of Ad Hoc Control procedure according to the formula

$$[TAR * CAR * (n_i + n_r)] * k > CACS \, , \qquad (8.2.4)$$

where k is activation coefficient,

n_i is a number of affected items on the moment of the i^{th} Ad-Hoc attempt,

n_r is a number of remaining items which have to be processed during delay of Active Control.

An additional check has to be performed if $n_r < PID/2$, then Active Control is suspended (called off) even if condition (8.2.4) holds. PID is a number of items which can be processed during delay of Active Control. For example, if 'Delay of effect' is 1 week, PID=230. Then if after 1 Ad-Hoc Control procedure 100 affected items are left, no Active Control is activated.

If the condition (8.2.4) is true, then Active Control will be applied. Otherwise the Ad-Hoc Control procedure is used. In case of applying the Ad-Hoc Control procedure, NC=0 and no SPOT properties are affected. In the case of Active Control, the effect has some delay effect. In this case the cost of dealing with risk will be:

$$Cost_{risk} = Cost_{AD\_HOC} + Cost_{AD\_HOC\_PID} + CACS\ ,$$

where $\text{Cost}_{\text{AD\_HOC}}$ is cost of Ad-Hoc Control procedure before activation of Active Control,

$\text{Cost}_{\text{AD\_HOC\_PID}}$ is a cost of Ad-Hoc Control procedure during delay before Active control has effect.

Active control will be activated for a given task only if sufficiently large number of items will pass through this task. From ORF’s experience, Active Control is a very rare event and up to now, Active Control was performed only once in the actual workflow (MXF Repair workflow) while more than 9,000 items were processed.

8.4.3 Classification of threats/risks in digital preservation

To classify possible threats in digital preservation, we have adopted the Simple Property-Oriented Threat Model (SPOT) for Risk Assessment. The 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
    - 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 workflows will fall in the six properties mentioned above, we introduce an extra possible state in the SPOT model for such cases: ***Other***.

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.

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 9, and described in Table 3.

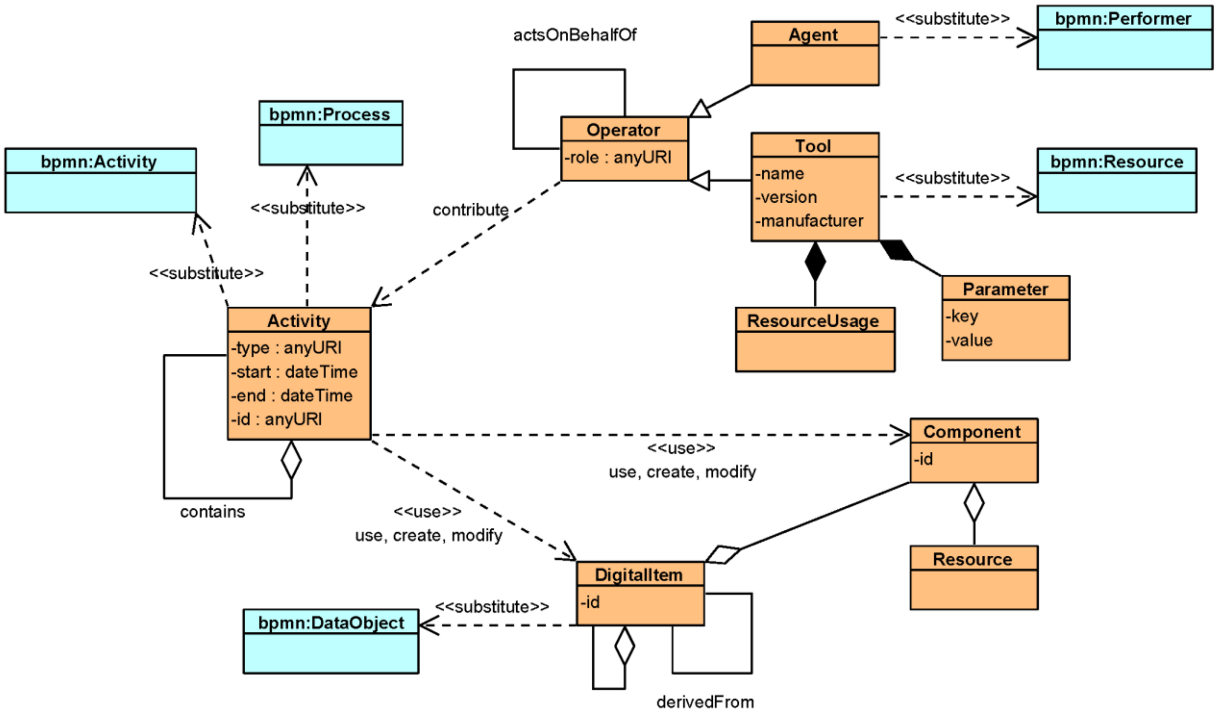


Figure 9: Entities of the preservation data model, their relations and the most important core properties. Blue entities are related BPMN entities.

Table 3: Description of the preservation data model 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*
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

Entity	Description	Relations
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 (in the context of a certain activity)
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.	

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.

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 documents what have 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[6] 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 9.

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

As many preservation processes are defined using BPMN [BPMN], it is important that information from workflows modelled in BPMN can be easily transferred to the preservation metadata model. Details about the relation of the preservation metadata model and BPMN have been addressed above. There are also other important data models, which we are considering here.

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]. Details about the interoperability are described in D3.1 and in [Bailer, 2014].

Another initiative considered in the design of the data model is the work in progress by MPEG to define the Multimedia Preservation Application Format (MP-AF). At the time of writing, the Draft International

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

Standard (DIS) of MP-AF is under preparation [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 and Tools

In order to qualify the Activity and Tool entities, controlled vocabularies have been defined. These vocabularies contain hierarchies of Activity and Tool types, which enable to better convey the semantics of the executed workflow. Due to the fact that they are modelled as hierarchies, a consumer of a preservation metadata document can also make use of the information, if it can only interpret the more general category of activities and tools but not a specific one referenced in the document. The proposed terms for specific Activities and Tools can be found in Annex C (Section 15).

Besides an XML representation of these controlled vocabularies, they are represented using the Simple Knowledge Organization System (SKOS) [SKOS] as well. SKOS defines a way to express a controlled vocabulary in RDF [RDF]. This framework enhances the machine-processability, interoperability, and reusability of controlled vocabularies.

In particular, one reason for describing the controlled vocabularies in terms of SKOS is to simplify interoperability with the risk management tools, which also represented preservation risks using ontologies (see Section 8.3). Another reason is that there are relations in such a hierarchy that should be captured as well. For example, a more general term can be determined when consulting the SKOS representation. This is useful if a tool encounters an unknown term in a metadata document. Furthermore, the SKOS representation supports links to additional existing ontologies. Thus provenance information described by W3C PROV ontology [PROV-O] or ontologies modelling BPMN entities can be linked together.

When using the SKOS approach for modelling a controlled vocabulary, all related terms are instances of class `skos:Concept`. In addition, SKOS provides properties in order to describe relations between these terms. For example, the property `skos:broader` is used to link a term to a more general one, whereas the property `skos:narrower` links to a more specific concept. Moreover, the entire classification scheme is classified as instance of `skos:ConceptScheme` and is described by appropriate metadata elements, such as creation and version information.

In Figure 10, an example for describing the terms and relations of a controlled vocabulary using SKOS is depicted. The exemplary terms are part of the controlled vocabularies for activities. First, these terms are described as instances of class `skos:Concept`. In order to establish a link from a term to a more general one, the propery skos:broader is applied. Thus this property is used to link from `ex:Checksum` to `ex:Checking`, and from `ex:Checking` to `ex:Activity`. The meaning of these relations is that `ex:Checking` is a more general term of `ex:Checksum`, while `ex:Activity` is the more general term of `ex:Checking`. Since this property skos:broader is defined as being transitive, it can be inferred that `ex:Activity` is a more general term of `ex:Checksum` as well.

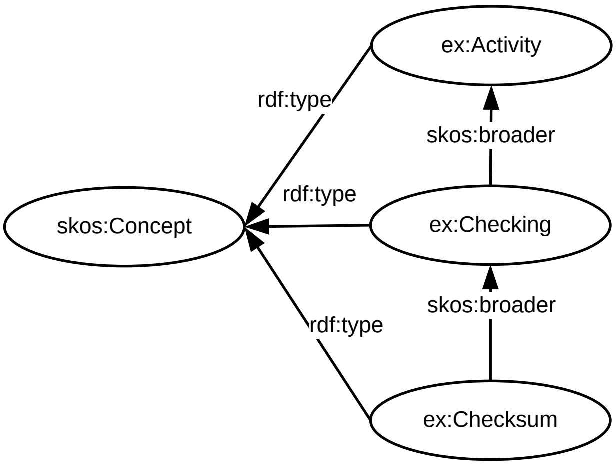


Figure 10: SKOS approach for describing a controlled vocabulary.

In contrast to this SKOS approach, activities and actions are modelled differently in the preservation risk ontology. Here, they are modelled in terms of classes and sub-classes. In contrast, in the SKOS approach all terms are instances of the class `skos:Concept` and relations between these instances are modelled by properties. However, the SKOS and OWL representation can be merged together according to the overlay pattern described in [OWLSKOS]. This pattern is depicted in Figure 11. Activities are modelled both as classes of the preservation risk ontology and as terms of the controlled vocabulary. These terms are described as instances of `skos:Concept`. In addition, a `rdfs:subClassOf` relation correspond to `skos:broader` relations. If needed, additional SKOS relations can be applied in order to describe relations between the activities in more detail. However, this overlay approach leads to a model requiring OWL Full, since the activities are classes of the preservation risk ontology and instances of `skos:Concept` at the same time. Therefore some caution is required, not to mix up these two different representations in use cases where OWL DL functionality is assumed. In such cases always one of the two representations should be excluded to simplify processing. This can easily been done while keeping the advantage of maintaining the set of activities only once.

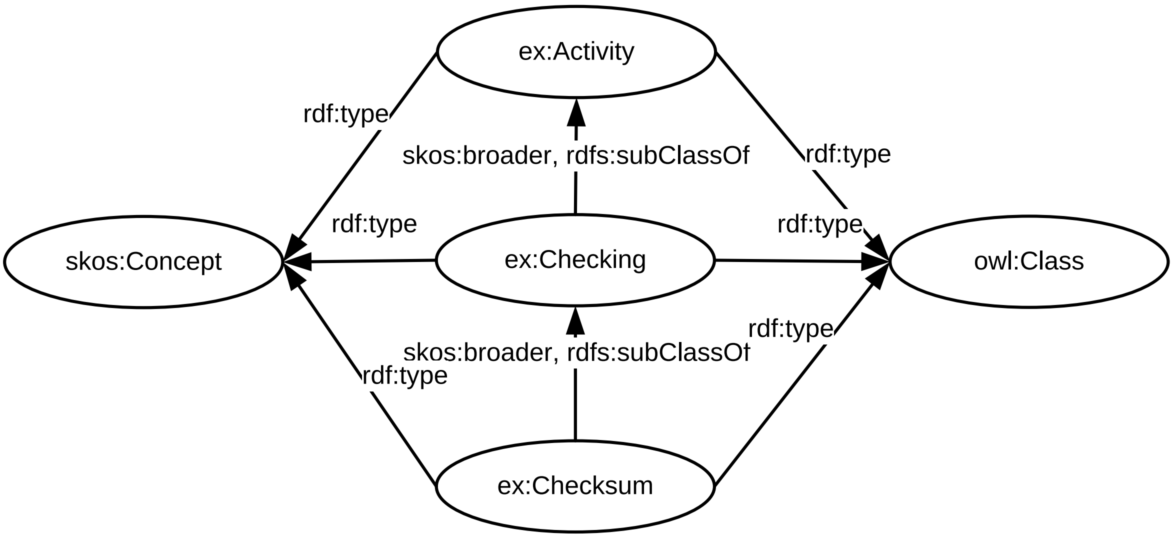


Figure 11: Overlaying pattern.

8.6 Preservation Metadata Service

8.6.1 Overview

A service for working with preservation metadata is currently being implemented. The service consists of three layers of components:

- Implementation of the entities of the data model
- Interfaces for ingesting source metadata (process descriptions, execution logs)
- Interfaces for accessing/serialising the preservation metadata

Figure 12 shows an overview of the service and its interfaces to other components. The core functionalities of the preservation metadata service are described in this section, the interface in the two following sections.

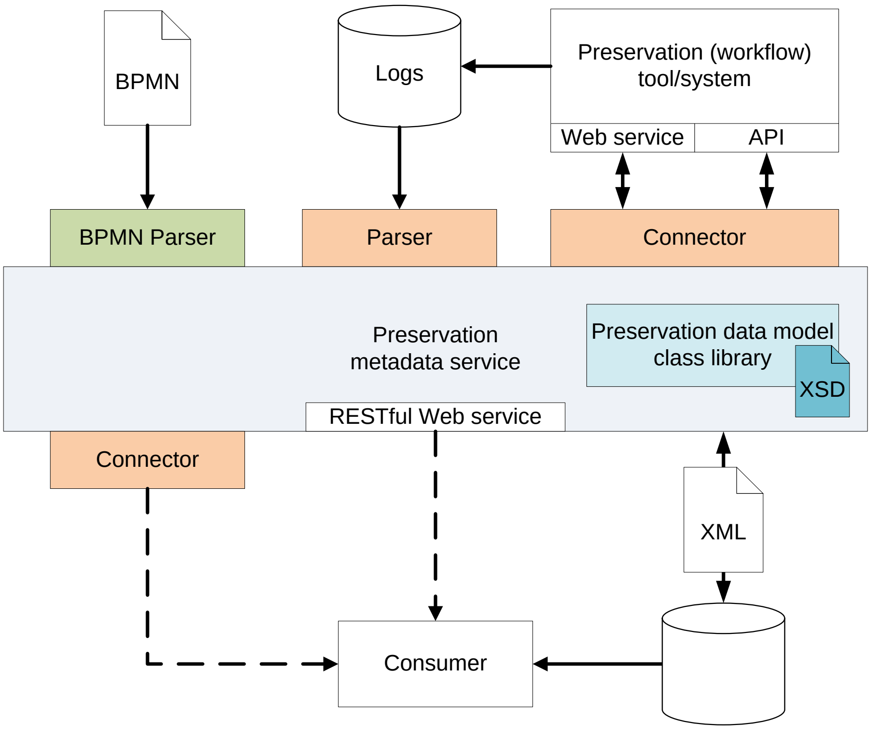

BPMN
Logs
Preservation (workflow) tool/system
Web service
API
BPMN Parser
Parser
Connector
Preservation metadata service
Preservation data model class library
XSD
RESTful Web service
Connector
XML
Consumer

Figure 12: Preservation Metadata Service and its Interfaces.

A C++ class library for the metadata model is generated from the XML schema representation of the model, using the code generation approach described in [Bailer, 2011]. This code generator was originally developed for handling MPEG-7. The preservation metadata model uses some other XML schema constructs, which require extension of the code generator.

The “static” information, i.e. the planned process and the interdependencies between services, can be directly derived from the BPMN model of the process. This results in a stub of a chain planned activities without detailed metadata. This information arrives from the data providers (see Section 8.6.2), and is added using the class library.

As workflow systems usually provide their own interfaces for live monitoring, it is for most applications sufficient to provide preservation metadata for completed process instances rather than performing live updates of running instances.

The collected preservation can be serialised to files, using the MPEG MP-AF format which is compatible to the defined preservation metadata model. A web service interface for accessing the data is described in Section 8.6.3.

8.6.2 Interface to Data Providers

The interfaces to data providers, i.e. workflow engines, services and tools, can be implemented via provided APIs or log files. In a first step, log files will be supported. As an example workflow, the MXF repair workflow deployed at ORF using Cube Workflow will be used.

Information can be collected on several levels. On the top level, a workflow engine or orchestration system provides information about the steps in the workflow being executed and the services being invoked, as well as their parameters. However, a service invocation on this level may in fact start a subprocess implemented in the service being called. In this case, more detailed information can be gathered from the tools and services implementing actual functionality on the next level. The information about the depth of this nesting comes from the BPMN file or a set of BPMN files describing the workflow.

A list of useful information to be gathered from data providers has been compiled. Implementations of connectors of log file parsers should strive to obtain this set of information, where possible (not all services will be able to provide the full set of information).

For the overall process, this includes the start/end times as well as users/operators invoking/monitoring the process. For each of the activities, the following information is interesting:

- start/end time of processing
- name/id of tool/service used, including version/manufacturer
- parameters/settings used and/or configurations/profiles of the tool/service
- name/id of content item(s) used/processed/created
- if applicable, metadata documents used/created
- users/operators interacting with the tool (type of task/input if available)
- runtime and resource consumption (CPUs/cores used, memory)
- result of the execution: success/failure, any errors occurring
- exception handling: calling alternative subprocess/service, notifying operator and the actions taken by operator, etc.

8.6.3 Interface to Consumers

The preservation metadata service also exposes an interface to any consumers who may want to access the collected preservation metadata. An example of a consumer of preservation metadata is the risk management framework, as discussed in Section 8.2 on page 33.

The metadata itself is represented in XML, following the MPEG Multimedia Preservation Application Format (MP-AF) standard (discussed above). A RESTful web service interface has been designed that allows three key operations: adding, retrieving and deleting preservation metadata documents.

For further details of the RESTful web service interface, please refer to Annex D (Section 16).

8.7 Rule-Based Decision Engine

The work on rule-based decision engines is a part of the workflow execution environment, depicted as Cube Workflow in Figure 2 on page 34 (Section 8.2.1). Research and Technology Development (RTD) on rule-based decision engines was originally planned for a specific use case on the cyclic migration of data tapes. However, this work has been generalised as it was clear early on that it is not domain specific, which is discussed below in Section 8.7.1. The following sections discuss the RTD work on this topic.

8.7.1 The intention behind the DAVID Rule-Based Decision Engine RTD

Typically, a data tape collection has to be migrated every few years, to the latest generation of tapes. For this migration, all data tape contents have to be copied to new data tapes. Migration from one LTO tape generation to a newer one is a critical step in data preservation and fraught with risk. Prior observations have indicated that up to now this infrequent non-trivial migration step has an increased operational risk and is mainly managed as a human centric project within the IT department. Process automation should provide the chance to establish a higher level of process control. Up to now there is not much tool support available for a controlled supervision of such cyclic migration step. Additionally, almost no experience from earlier migration can be used, as the frequency is too low to establish best practice procedures based on own experiences. Tool support with business process modelling and integrated rule-based decision engines has the potential to help improving the robustness of this process as well as provide machine readable documentation (automated process logging) that can be used in predictive analytic models to update risk analysis estimates for future migration steps.

With both DAVID archive partners, INA and ORF, the former data tape migration strategy was discussed. It was decided early to generalise the focus of the rule engine RTD to a wider scope as the technology behind a rule-based decision engine is not domain specific to data tape workflows.

8.7.2 *Decision Modelling - From Isolated Decisions to Decision Engines*

Decisions can be strategic, tactical or operational. Typically, the number of decisions expands the same way from a few on the strategic level to quite a lot at the operations side. The economic impact of every decision shows the opposite direction.

The Digital Disruption leads to rapidly changing business environments and generates a fast growing pressure to maximise efficiency in the media industry by establishing fabric approaches and foster process automation wherever possible. This leads straight to business process management (BPM) systems and the paradigm of Service Orientated Architectures (SOA). The decision-making in Decision Management Systems is dynamic and change is to be expected. The way a decision is made must be continually challenged and re-assessed so that an organisation can learn what works and adapt to work better. A next logical step is to establish Decision Management Systems to separate business logic from business processes. Through this separation of concerns the business decision logic can be maintained and updated independently.

8.7.3 *Business Process Improvement through Decision Automation*

Advances in technologies have enabled process automation to perform many traditional business workflows more efficiently, reliably and economically than human-driven work can achieve. However, humans are still imperative in the automated workflow management system due to their decision-making functions. Current automated systems lack important abilities of humans like ingenuity, creativity, and reasoning. For this decision automation systems may fail seriously under most circumstances which were not anticipated in beforehand. When a human acts as a supervisor or a monitor of an automated system, he/she has to maintain a clear situation awareness to make an informed decision. To implement such process-aware (self-awareness) approaches to an automation system is still a challenging RTD task for rule-based decision engine software designers.

8.7.4 *Requirements for a Generalised Rule-Based Decision Engine*

In a generalised approach to solve a decision problem, decision rules may be applied as part of a workflow management system (WfMS). A decision rule corresponds to a mathematical function that, given a set of parameters, computes the best alternative based on the criteria important for a particular decision. Therefore, integrating such decision rules into business process models can make decisions in a process more explicit and allow process participants to select the best alternative at a certain decision-making point.

In automated processes, parameters to drive decisions can be described as events. Events can be provided from external sources. Events can be extracted information from the business process itself, but it can also be metadata which is extracted or calculated by media analysing tools from the media within the process. In the rules-with-actions approach a rule-based decision engine allows a business process to react to single or combinations of such events. The action is to branch the process depending on the value of the event. In a BPMN process execution framework, a decision engine can be used to control the token flow for branching control flow gateways (XOR-Split).

An event-driven WfMS with a rule-based decision engine fosters the explicit formulation of rules. By encapsulating rules in rule-engines, the desirable separation between rules and processes is achieved.

Desirable features for the generalised rule-based decision engine are:

- hide (encapsulating) complexity
- syntax checking
- semantic validation
- enabling business-IT-alignment
- scalability
- rule update in active process instances
- support for hard and soft constraints
- support for decisions based on unreliable input data
- adaptive and context-aware process execution

Some of these features will be addressed in more detail within the following paragraphs.

8.7.5 Implementation Variants for Rule-Based Decision Engines

Traditionally rules are documented in system specifications in either pseudo-code or natural language statements. Pseudo-code has the advantage of being unambiguous but is often unintelligible to business stakeholders. Natural language, on the other hand, has the advantage, if properly used, of being intelligible to business stakeholders. However, natural language can be ambiguous. Over a long period of time, methods to implement decisions in a flexible way have improved.

Common Concepts

To start with basic concepts: **decision trees** are a common concept for classification of known classes to a given set of data. Typically, decision trees are binary trees and each node corresponds to one Boolean question. The classification starts in the root and the results can be seen in the leaves. During the traversal of the tree, one path through the nodes according to the given input data is taken, guiding to the answer leaf. In practice, decision trees fit best for applications with a small set of around 15 outcomes.**Decision tables** are another common concept for making decisions in software systems. Each column denotes one input, each row one output. The cells hold the conditions for a specific outcome. Such tables can handle a larger number of outcomes and can work with complex combinations of a few input conditions.

Scripted Decisions

Scripted decisions are a fast way to implement conditional flows in a BPMN process engine. For each flow a short script can be set up, which accesses only the required metadata and determines if this flow has to be taken or not. If a dynamic scripting language is used, changes can happen not only in design but also with running process instances (at run-time). A disadvantage of scripted decisions in standard imperative programming languages is the missing decoupling of application code and rules. For decision modelling, script code is used to ease the coding (e.g., Groovy Script for Java Enterprise Environments for efficient byte code production). Maintaining complex scripted decisions over time can become error prone and the code can look confusing, especially for business people.

Rule Engine based on Declarative Approaches

Declarative programming is a programming paradigm that expresses the logic of a computation system without describing its control flow. This is in contrast to an imperative programming style where explicit steps are described. Declarative programming uses a higher level of abstractions and often considers computations as deductions in a formal logic space.

This way decision logic (business rules) can be documented separately to programming code in a pseudo natural language. Rules can be updated during process execution (at run-time). Disadvantages are the higher complexity and significant demand on computer RAM to cache pre-calculated terms to speed up the decision making process.

GUI based interfaces

GUI based interfaces support the maintenance of decision rules by a self-service interface, directly usable by process analysts and business users responsible for the process operations. This can improve agility and the continuous quality improvement process within an organisation. The chosen GUI concept is widely independent from the used technology in the decision engine back end.

To evaluate the practical user acceptance, Cube-Tec has tested different approaches to see whether a user interface has accomplished to offer an easy enough view on the decision to be made and the ability for maintenance.

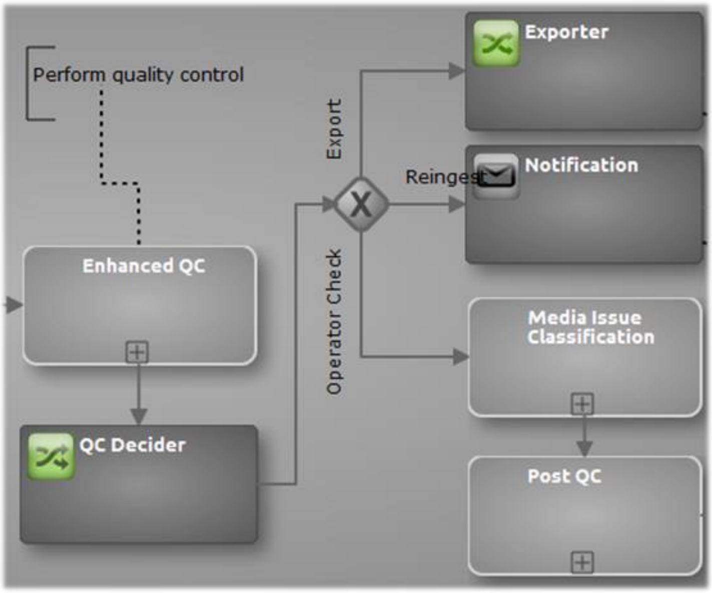


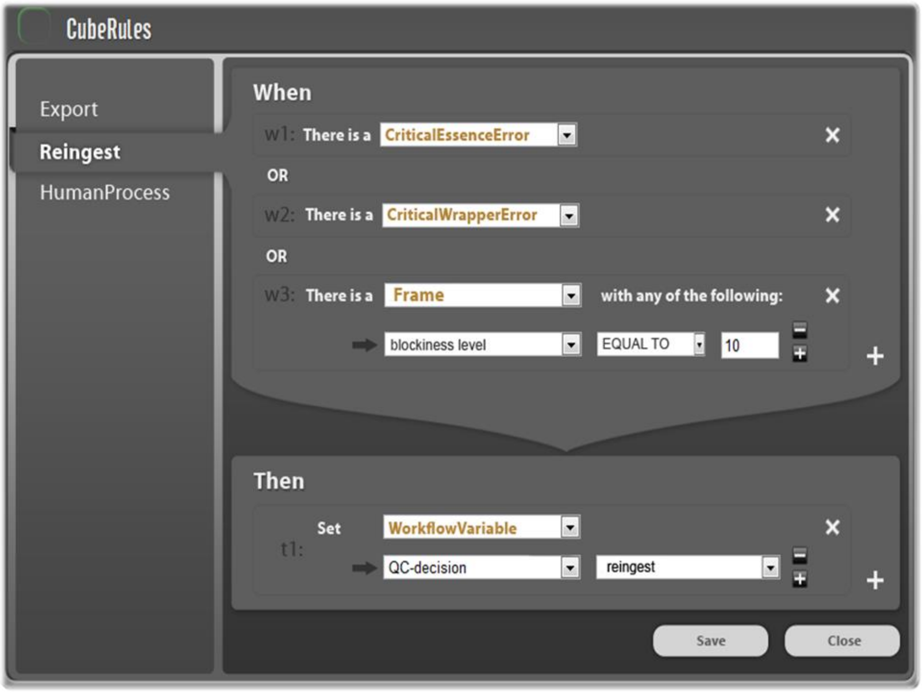
Figure 13: Example BPMN process within a QC process with a decision point (exclusive or gateway).

Figure 13, above, shows an example BPMN process within a QC process to model the decision of an exclusive gateway which decides whether a human operator has to be involved to double check QC results, whether the QC indicates massive errors that require a redo of the earlier process steps, or whether automatic processes can directly move on with the next steps.

Cube-Tec has set up a demonstrator to gather practical experience and user feedback in understanding and designing rules via a web browser with a dynamically adaptable user interface – see Figure 14 below. The rules were organised as:

“**When** `w1` or `w2` or … **Then** set variable `t1` and `t2` and …

(“+” and “-” buttons in the decision structure can be used to expand the structure)



When

w1: There is a CriticalEssenceError

OR

w2: There is a CriticalWrapperError

OR

w3: There is a Frame with any of the following:

blockiness level EQUAL TO 10

Then

t1: Set WorkflowVariable

QC-decision reingest

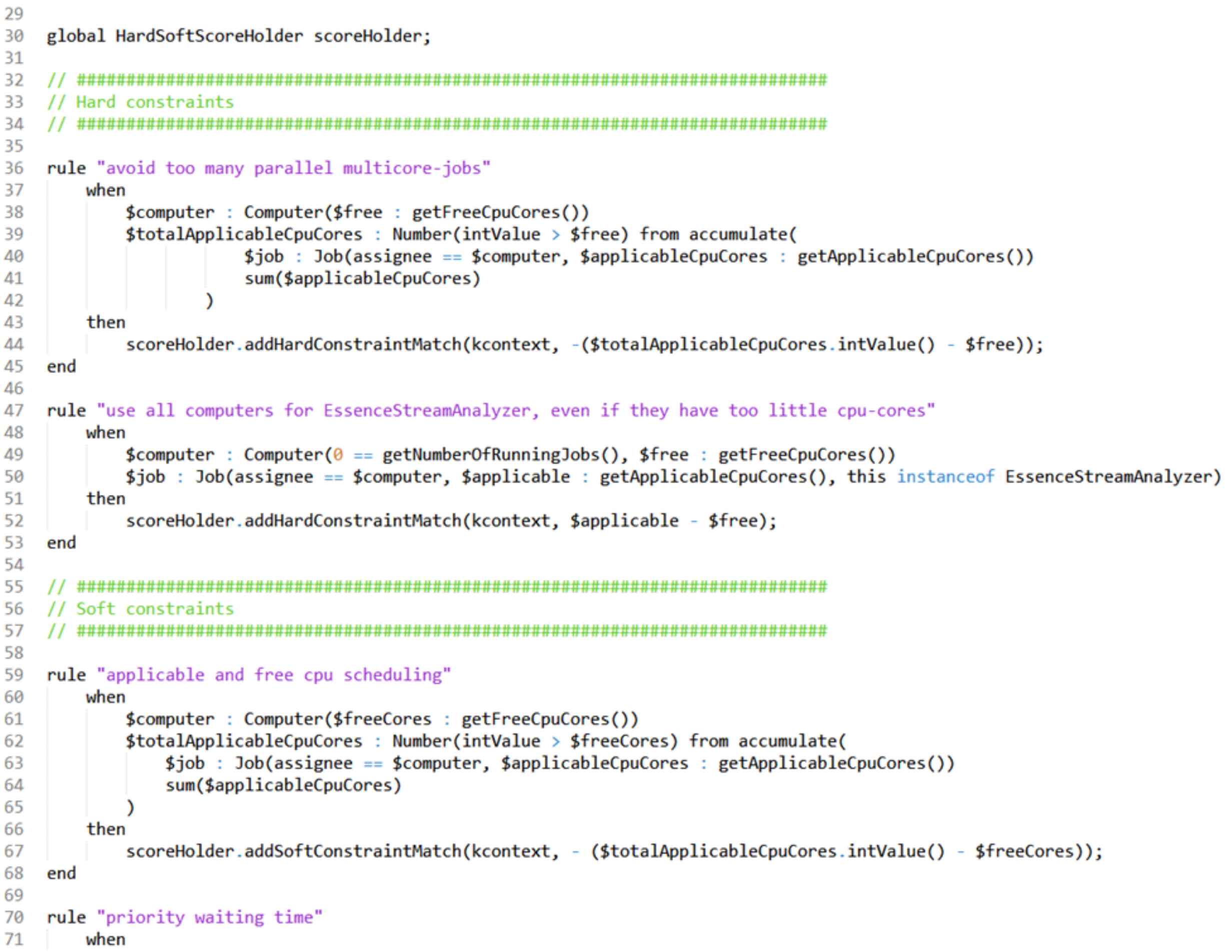
Figure 14: Example GUI front-end for a rule-based engine to evaluate user acceptance and user preferences.

8.7.6 Practical Use Cases for Declarative Rule-Based Engines

Event enrichment and BPMN workflow branching is the originally planned-for use case for the rule-based decision engine in the DAVID risk based framework. This is realised highly efficient within the format compatibility tools (commercially named MXF Legalizer). In addition practice has shown the power of these engines within dynamic optimisation tasks for process resource scheduling tasks in flexible environments. So, another use case for the DAVID risk management framework is the dynamic load balancing of automatic BPMN tasks. To maximise in every moment the usage of the available hardware resources within a multi-core and multi-CPU system especially in a distributed server farm requires rule engines which manage these dynamic constraints efficiently. The Cube-Tec implementation has shown that a declarative rule-based engine is well suited for this application type.

This enables at every moment an optimised utilisation of available hardware resources. In the use case of robust and reliable server farms this enables also the dynamic fail-over management in the case that single components have to be replaced during run-time in the redundant deployment environment.

Figure 15, below, shows a small crop of a declarative definition of scheduling rules for a load balancing simulator for hardware resources.



```

global HardSoftScoreHolder scoreHolder;

// #############################################################################
// Hard constraints
// #############################################################################

rule "avoid too many parallel multicore-jobs"
    when
        $computer : Computer($free : getFreeCpuCores())
        $totalApplicableCpuCores : Number(intValue > $free) from accumulate(
                    $job : Job(assignee == $computer, $applicableCpuCores : getApplicableCpuCores())
                    sum($applicableCpuCores)
                )
    then
        scoreHolder.addHardConstraintMatch(kcontext, -($totalApplicableCpuCores.intValue() - $free));
end

rule "use all computers for EssenceStreamAnalyzer, even if they have too little cpu-cores"
    when
        $computer : Computer(0 == getNumberOfRunningJobs(), $free : getFreeCpuCores())
        $job : Job(assignee == $computer, $applicable : getApplicableCpuCores(), this instanceof EssenceStreamAnalyzer)
    then
        scoreHolder.addHardConstraintMatch(kcontext, $applicable - $free);
end

// #############################################################################
// Soft constraints
// #############################################################################

rule "applicable and free cpu scheduling"
    when
        $computer : Computer($freeCores : getFreeCpuCores())
        $totalApplicableCpuCores : Number(intValue > $freeCores) from accumulate(
            $job : Job(assignee == $computer, $applicableCpuCores : getApplicableCpuCores())
            sum($applicableCpuCores)
        )
    then
        scoreHolder.addSoftConstraintMatch(kcontext, - ($totalApplicableCpuCores.intValue() - $freeCores));
end

rule "priority waiting time"
    when
```

Figure 15: Code snippet of a declarative definition of scheduling rules.

8.7.7 *OMG Decision Model and Notation (DMN) - the New Standard*

In February 2014, the Object Management Group (OMG) has released a first public beta release of a new standard on the topic of decisions: Decision Model and Notation (DMN). It is supposed to close the gap between decision design and decision implementation. On one side, it offers a standardised notation for complex decisions and on the other, it provides a data model for representing those decisions in an interoperable way. The field of application ranges from simple decision, with human operators as authorities for those decisions, to completely automated decision services. Applying DMN to a workflow engine context would result in several benefits. The graphical notation language is easy to understand – for IT staff, business architects and non-IT business users – and integrates seamlessly with BPMN. Modifications can be made during run-time within a process instance and in a collaborative manner depending on user roles and access rights and can be controlled by a corporate rights management system. In addition software designers are able to make the customer see only the part of the decision which is expected to be changed. This helps to keep complexity for the user manageable (under control). Using DMN this way massively cuts the amount of metadata a user has to deal with and helps avoiding mistakes. To achieve this advantage, more effort has to be put in the development of a certain workflow and its decisions.

DMN integrated notation for decision management in very much the same way BPMN does for business processes. DMN enables modellers to address business modelling from rule-oriented, process-oriented, and information-oriented views of businesses.

8.7.8 *Conclusion and Outlook*

Different techniques can be used to realise rule-based decision engines, each with its strengths and drawbacks. Imperative script-based as well as declarative versions are now implemented in productive prototypes within the Cube-Tec format compatibility tools and in the Cube Workflow framework.

To reach the next level in automating business processes, decision engines need a better foundation to perform not only complex but also intellectually challenging tasks with decision under uncertainty. The human decision-making mechanism is traditionally modelled in a BDI (belief, desire, and intention) paradigm. There are great models in use to improve BDI models for decision automation.

The last years have shown an impressive progress of interdisciplinary research in domains like: computational intelligence and goal-oriented reasoning, probabilistic inference or more generalised soft computing and machine learning techniques. Those can be supported by well-founded, stable semantics with RDF and OWL ontologies and are preparing themselves to enter the public stage for the next generation of decision engines.

9 Analysis of ORF Risk Management

Following the description of the risk simulation modelling work in Section 8.4 (above on page 40), we show in this section how the model can be applied to a real-world workflow; the ORF MXF Repair workflow.

9.1 ORF MXF Repair Workflow

The ORF MXF Repair workflow can be seen below in Figure 16. This workflow was selected due to its simplicity for testing purposes.

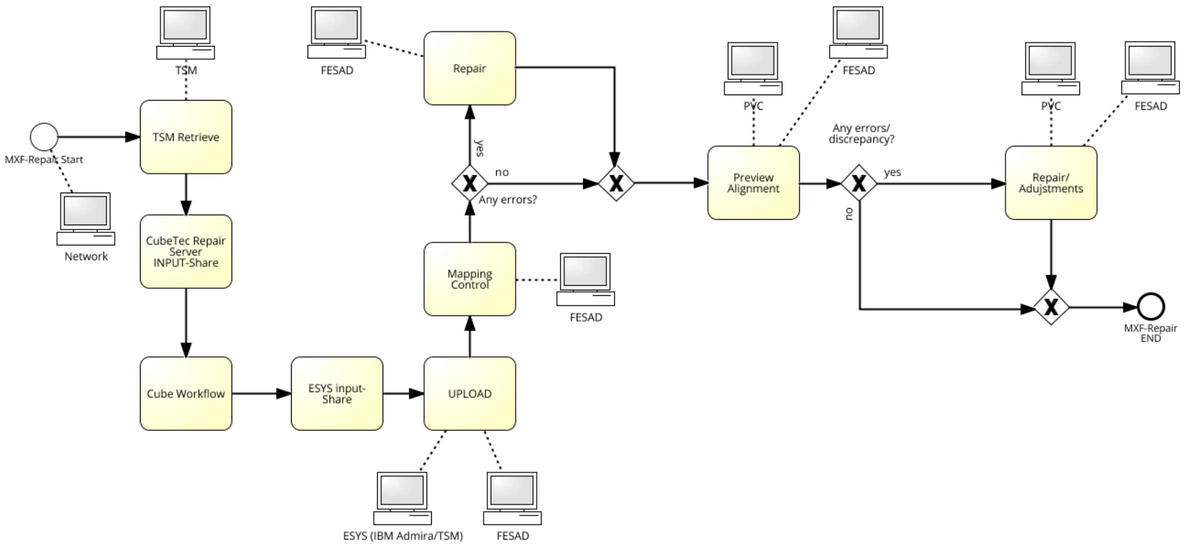


Figure 16: ORF MXF Repair workflow.

The ORF MXF Repair workflow consists of 9 tasks and two exclusive XOR conditions. At this stage IT systems were not simulated in our model. Also, note that 'Cube Workflow' is treated as a black box in the ORF MXF Repair workflow, but is a workflow in itself that is executed by Cube Tec in their workflow execution software of the same name as the task name.

We cannot disclose details of all the risks and the related information such as financial impact, etc., for this workflow. However, Annex B (Section 14 on page 79) gives information on risk specification for 2 tasks. Below we give some general information on the risk modelling for this workflow before presenting the simulation results in the following sub-section.

In this workflow, each task has got two risks that can occur. For some tasks, the two risks are mutually exclusive, i.e., they cannot occur simultaneously; e.g., for tasks 'TSM Retrieve', 'Cube-Tec Repair Server INPUT-Share' and 'ESYS input-Share'. The task 'Upload' has got two risks that can either happen independently or simultaneously, so a combined frequency of two risks occurrence is provided for this task. Moreover, if the risk 'Copy Error' takes place in the previous task ('ESYS input-share'), then the frequency of the risk 'Fails' (for the 'Upload' task) increases slightly (temporally). For the rest of tasks, risks have the following dependencies: initially the risks are modelled as mutually exclusive, however if the 'Overload' risk takes place then the frequency of the second risk increases by the given Multiplication factor and occurrence of this second risk is simulated with the new frequency. That is, if 'Overload' occurred, it can cause the second risk such as wrong assessment or wrong mapping to take place too. See Annex B (Section 14 on page 79) for the configuration details for this.

After the task ‘Mapping Control’, the simulation procedure will check whether any errors took place (the first XOR gate). If there are errors, then the ‘Repair’ task is performed. The probability of any errors is 0.1%, i.e., the ‘Repair’ task will take place very seldom during the workflow runs. A condition after the ‘Preview Alignment’ task checks whether any error or discrepancy is present. It is known that probability of this error/discrepancy is 60%.

The Estimated Frequencies for each risk are estimated based on the annual throughput of approximately 12,000 items (1000 monthly, 230 weekly, 46 daily). It is assumed that the ORF MXF Repair workflow runs 15 working hour per day, 5 days a week, 22 working days in a month and 264 working days in a year.

The risk Simulation tool was developed in Matlab (using the Simulink toolbox) [Matlab] and is used for running scenarios for ORF MXF Repair workflow. The scenario reported here is processing 1,000 items to analyse the risk occurrences and their consequences, as well as the usage of Ad-Hoc Control procedures. The simulation results for ORF MXF Repair workflow are presented in the next subsection.

9.2 Simulation Results

First of all, we will analyse how effective the Ad Hoc Control procedure is at dealing with risk. For this purpose we consider a scenario when 1,000 items processed via a workflow 120 times, which is equivalent to 10 years running of the one ORF MXF Repair workflow or 120,000 items processed via the workflow.

In Figure 17, the top graph shows how many times different risks took place and how many risks were undetected overall during 120 runs, while the bottom graph shows a number of affected items during this period. From this we can see that ‘Preview Alignment’ had the most risk occurrence during the 120 runs, followed by ‘Upload’ and ‘Mapping Control’. ‘TSM Retrieve’ and ‘Cube-Tec Repair Server INPUT-Share’ have the smallest number of risk occurrences. No risks took place for the ‘Repair’ task, since it is a rare event and it was not evoked even a single time during 120 runs. Some risks will be undetected for all tasks.

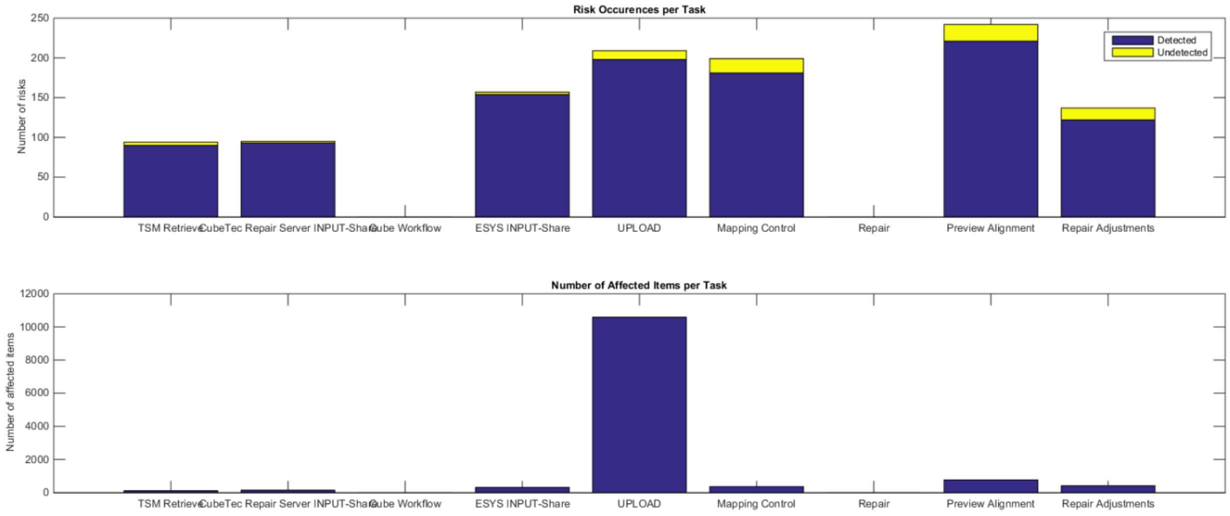


Figure 17: Risk Occurrence and Number of Affected Items.

However, the risks occurring during the ‘Upload’ task will affect the largest number of items (in excess of 10,000 items per 120 runs, or, on average, around 90 items per run) – see the bottom graph in Figure 17. The number of affected items during ‘Preview Alignment’ was 768 for 120 runs (on average 3 items per run), which followed by 421 affected items during ‘Repair Adjustment’. The least number of affected items were during ‘TSM Retrieve’ and ‘Cube-Tec Repair Server INPUT-Share’ tasks (between 0 and 3 affected items per run). Note that even though the risks occurrence are very frequent for the ‘Mapping Control’ task, the number of affected items is low (around 2-3 on average).

Next, we investigate how the Ad-Hoc Control procedure and Active Control cope with these risks. Figure 18 shows a sum of Negative Consequences (NCs) for each task with Ad-Hoc Control (bottom graph, NC per Task with Control) and without (top graph). If no Ad-Hoc Control procedure is in a place, then the largest NC from risks will be for ‘Mapping Control’, ‘Preview Alignment’ and ‘Cube-Tec Repair Server INPUT-Share’. The Ad-Hoc Control procedure reduces NC to zero for all risks that have been detected. After the Ad-Hoc Control procedure, NC are reduced practically by an order and the tasks with largest NC are ‘Mapping control’, ‘Preview Alignment’ and ‘Upload’ due to risk detection rates at these tasks. Note that ‘Preview Adjustment’ does not have any NC after the Ad-Hoc Control procedure, though the risks occurring here do not have 100% Detection level (see Figure 17 top). It is due to the fact that the risk ‘Overload’ has some NC and 100% Detection Level, while risk ‘Wrong Assessment’ though has Detection Level 75%, does not have any NC according to the provided data. See Annex B (Section 14 on page 79) for the configuration details for this.

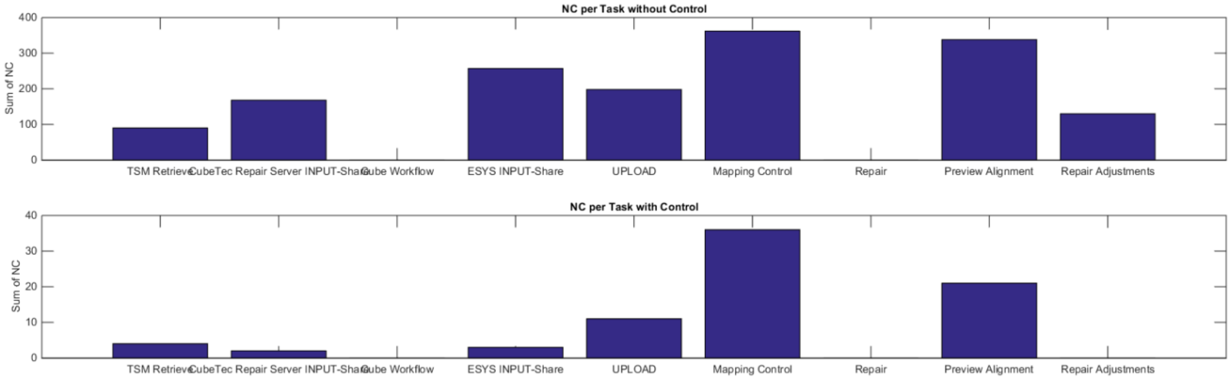


Figure 18: Negative Consequences with and without Ad-Hoc Control.

SPOT model impact can be seen in Figure 19 both without Ad-Hoc Control procedure (top graph) and with Ad-Hoc Control procedure (bottom graph). Again, the Ad-Hoc Control procedure reduces all effects on SPOT model properties for the detected risks. After the Ad-Hoc Control procedure, the most affected SPOT properties are Authenticity, Identity and Availability. Recall that we present here results based on overall 120 Monte Carlo runs, that is if we are interested in only processing 1,000 items once, then after Ad-Hoc Control procedure it is very likely that we will not observe neither any NC nor any affected SPOT properties for this workflow scenario.

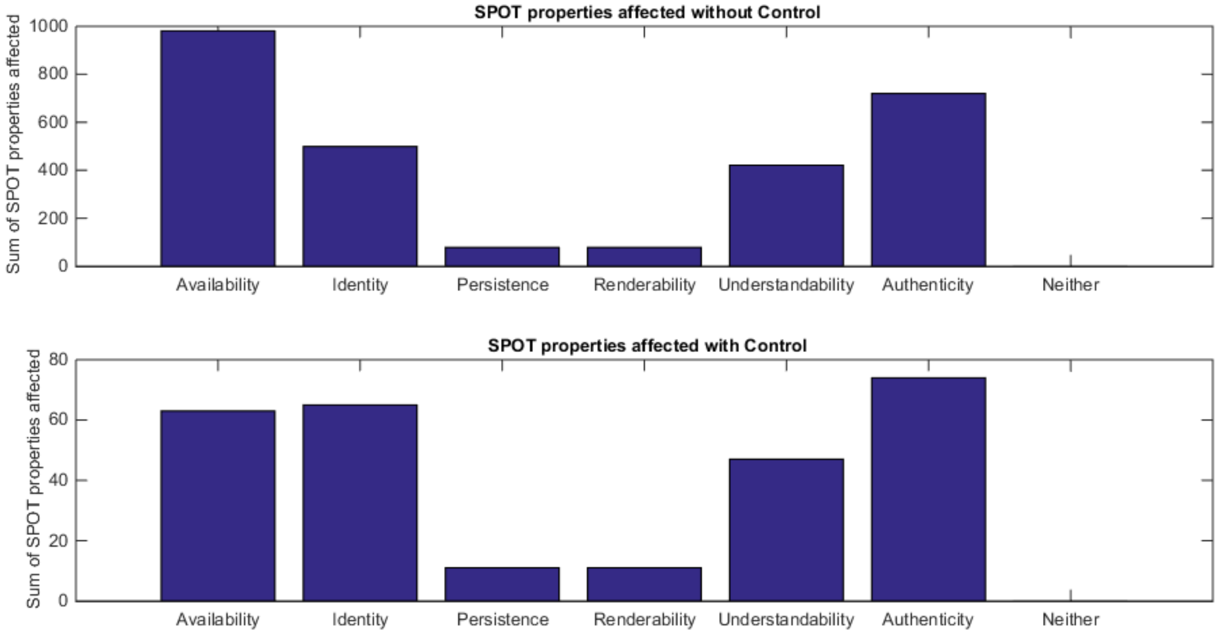


Figure 19: SPOT properties affected by risks.

The Ad-Hoc Control procedure is very successful for ORF MXF Repair workflow in dealing with risks. Figure 20 shows how often Ad Hoc Control was used for each task during the 120 runs (top graph) and the overall cost (bottom graph). Although Ad-Hoc Control was used for all task, it incurred additional costs only for ‘Mapping Control’, ‘Preview Alignment’ and ‘Repair Adjustment’. Risks occurring during these tasks incur the highest expense during the procedure for dealing with risks. It is due to the fact that, in the remaining tasks, Ad-Hoc Control is covered by overheads.

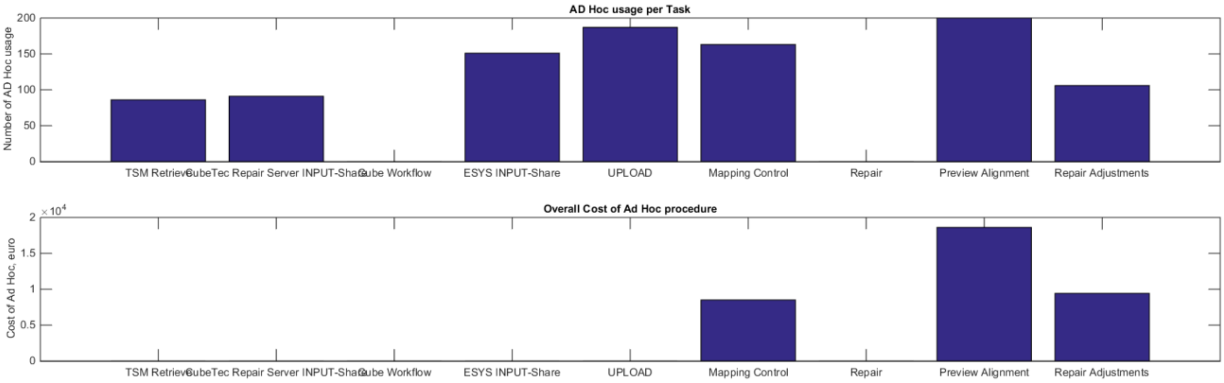


Figure 20: Ad-Hoc Control procedure usage and its overall cost.

Active Control was only activated once, during the ‘Repair Adjustment’ task. This was for the risk ‘Overload’, which resulted in a cost of €500 in this simulation scenario.

So far we did not consider individual risks. Although the Ad-Hoc Control procedure of dealing with risks is very effective in this particular workflow, it is of interest to rank risks according to impact on a workflow based on some score in the general case. Table 4 shows the ranking of risks according to impact score, which was calculated as product of affected items and negative consequences based on 120 runs of the ORF MXF Repair workflow under condition that no Ad-Hoc Control procedure was in place. The ranking allows us to identify the risks with the highest impact on the workflow. In case of the ORF MXF Repair workflow, the risk with the highest impact is `Fails’ during task ‘Upload. Although level of severity is not very high for this risk, it can occur several times per day. Other scores or criteria can be used to rank risks, such as according to additional cost of dealing with risks.

Table 4: Risk ranks in descending order.

Rank	Risk	Task	Impact	Score
1	'Fails'	'Upload'	Very High	10462
2	'Overload`	'Preview Alignment'	Very High	1042
3	'Overload`	'Repair Adjustments'	High	560
4	'Overload`	'ESYS INPUT-Share'	High	486
4	'Overload`	'Mapping Control'	High	486
5	'Overload`	'Cube-Tec Repair Server INPUT-Share'	Medium	252
6	'Wrong Assessment'	'Preview Alignment'	Medium	245
7	'Wrong Assessment'	'Mapping Control'	Medium	228
8	'Wrong Parameters'	'Upload'	Medium	127
9	'Copy Error'	'ESYS INPUT-Share'	Low	70
10	'Wrong File selected'	'TSM Retrieve'	Low	60
11	'Retrieve Fails'	'TSM Retrieve'	Low	55
12	'Copy Error'	'Cube-Tec Repair Server INPUT-Share'	Low	19
13	'Wrong Assessment'	Repair Adjustment	Very low	0
13	`Wrong mapping'	'Repair'	Very low	0
13	'Overload`	'Repair'	Very low	0
13	'None'	'Cube Workflow'	Very low	0

9.3 Evaluation of Simulation Results by ORF

Based on the results discussed above, this section is an evaluation from ORF based on the observations they have made in reality. Both ranking of Risk/Task and the impact is giving an exact picture of the actual situation ORF experienced when running the MXF-Repair. For example, they had several upload-fails per day and 'Overload' was indeed the greatest issue in 'Preview-Alignment'.

Even the ranking (Table 4, above) is nearly identical to ORF's experience; only the 'Overload ESYS Input Share' is actually a bit higher than in the results from the simulations. However, this is due to external reasons; the Upload-Share is used by several other workflows and a weak coordination between those and the MXF-Repair workflow caused these additional overloads, which, therefore, should be neglected. See Table 5 for a comparison of the ranking from the simulation results compared with the ORF ranking.

Also at the very end of the ranking the results match ORF's experience, which impressed their workflow-designers very much. Also, in this section of the results, there is only one swap in ranking positions. In the actual workflow 'Retrieve Fails' was slightly higher than 'Wrong File selected'. This may be due to the fact that in the early part of the MXF Repair process they experienced some severe network-issues.

Table 5: Risk ranks compared to actual (ORF) Rank.

Rank	Risk	Task	ORF-Rank
1	'Fails'	'UPLOAD'	1
2	'Overload`	'Preview Alignment'	2
3	'Overload`	'Repair Adjustments'	4
4	'Overload`	'ESYS INPUT-Share'	3
4	'Overload`	'Mapping Control'	5
5	'Overload`	'Cube-Tec Repair Server INPUT-Share'	6

Rank	Risk	Task	ORF-Rank
6	‘Wrong Assessment’	'Preview Alignment'	7
7	‘Wrong Assessment’	'Mapping Control'	7
8	'Wrong Parameters'	'UPLOAD'	8
9	'Copy Error'	'ESYS INPUT-Share'	9
10	'Wrong File selected'	'TSM Retrieve'	11
11	'Retrieve Fails'	'TSM Retrieve'	10
12	'Copy Error'	'Cube-Tec Repair Server INPUT-Share'	12
13	‘Wrong Assessment’	Repair Adjustment	13
13	`Wrong mapping'	‘Repair’	14
13	'Overload`	'Repair'	14
13	'None'	'Cube Workflow'	x

The results for Risk Occurrence and the Number of affected Items show a similar picture; the results of the simulation match with our experience during the actual workflow. Just the absolute number of affected items for Upload-Fails seems to be too high; a first explanation may be the fact that in the actual workflow those Upload-Fails had a significant concentration in the first 6 months of the MXF-Repairs (due to several changes in the affected ESYS-systems) – and from the same timeframe the values for the simulation-model has been taken.

The effect of Ad-Hoc Control, shown in Figure 18 (above on page 63), is a very strong argument and help for implementing a proper instalment for control measurements. And again the figures for NC in the different tasks reflect very well the actual situation in the MXF Repair workflow; very interesting is the rather high effect of “technical” measurements (e.g. in ESYS Input Share) compared to those with “human-related” measurements (e.g. Mapping Control).

The high impact of Ad-Hoc Control reflects again the actual experience and strengthens our arguments in “investing” in those by having a rather large “Overhead” available. Especially the results in the SPOT model for Ad-Hoc Control are very impressive and promising. Finally the results for costs will be a real asset-argument for the next budget-discussion on future workflows in the domain: the extremely positive impact of Ad-Hoc Control was always neglected or doubted by the decision-makers here; but the results from this analysis will really be hard to be ignored!

It has to be stressed here, that the quality of the results are based not only on the model, but also on the quality of the input-data given. ORF put very much effort and time in providing accurate and reliable data for both the model and the simulation, so it has to be expected that the old archive-rule “Garbage in = Garbage out” is valid for risk management and assessment as well. You have to invest a good percentage (in case of the MXF-Repair workflow approx. 1/3) of the budget reserved for “accompanying measures” for this task and the calculation and survey of proper planning-data to get good results and actually have the chance to save money in the actual workflow or process. And being an expert or involving experts in the domain is highly necessary to save efforts in this process of planning and data-collection.

For all colleagues at ORF involved in the process, the tool proves to be a very good and highly reliable instrument to evaluate risks and their actual impact even before or at an early stage of the implementation of a workflow. However, as discussed above, if the model is not configured with the right (and accurate) data, the results will not match with reality. Therefore, it is important to complete the risk management process, based on the Deming cycle, to capture monitoring information from the workflow executions to update both the risk information and simulation configuration to improve the accuracy. Refer to 8.2.2 on page 35 for the discussion on this.

10 Conclusions

This report has defined the scope of digital damage for the DAVID project and proposed a conceptual risk framework that addresses the need for long-term quality assurance of digital assets. Firstly, the problems affecting digital Audio-Visual (AV) content has been discussed, and we define the term 'digital damage' as:

> *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, transferring or accessing the content.*

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 other approaches can be modified from previous experience. However, the effectiveness of existing 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, which has been explored in more depth in D2.2 [Hall-May 2014], 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 the 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 due to systematic errors such as format choices, tool misconfiguration and process changes.

In this deliverable, we have proposed a risk management framework aimed at assisting preservation experts design more robust workflows and optimise existing workflows / change processes. In the context of the topics addressed in DAVID, this risk management framework is key to the prevention of digital damage, aiming to help organisations move away from "firefighting" – i.e., organisations may spend more time dealing with issues rather than preventing them in the first place.

The proposed risk framework covers aspects of designing workflows, specifying risks for workflow activities/tasks and how they are controlled, simulating workflow executions and analysing preservation metadata from the real workflow executions in order to further improve the workflows and update the information on risk occurrences and other simulation parameters. We have aligned the use of the risk framework with a best practice process based on the Deming cycle (also known as the PDCA cycle), as well as showing how this relates to the ISO 31000 risk methodology. The Deming cycle is a four-step iterative method commonly used for control and continuous improvement of processes and products, and is key to, for example, ITIL Continual Service Improvement [Lloyd 2011]. In general terms, risk management is a part of continual improvement of processes – preservation workflows in this context.

We have presented a novel preservation risk ontology, which has been designed to encapsulate domain knowledge generated in the DAVID project about known risks and controls for preservation activities. Based on this ontology, semantic reasoning is used to a) determine whether risks are mitigated or not (depending on controls available) and b) to assist preservation experts in the risk specification by proposing risks to the activities in the workflows that are being designed.

A risk simulation model has also been presented in this report, allowing workflow executions to be simulated in order to analyse the effects of deploying risk treatment strategies. A key aim of this is to improve cost-benefit by a) identifying and understanding key vulnerabilities and b) targeting investments to address those vulnerabilities. We have presented results of applying the simulation modelling on a real-life workflow, the ORF MXF Repair workflow. The simulation results are very positive, almost 100% matching the observations ORF have made from the real executions of the workflow with respect to both the ranking of risks and their impact.

A key part of the risk management process is to utilise the information from the workflow executions to enable continual improvements (completing the Deming cycle mentioned above). A preservation metadata model and service has been presented in this report, which provides evidence to support and improve the risk assessment approach. The use of rule-based decision engines have also been investigated, analysing the pros and cons of different techniques that can be used to further advance the automation and optimisation of the decision making in business processes. Both script-based and declarative interfaces have already been implemented and integrated into Cube-Tec’s format compatibility tools and in the Cube Workflow framework.

A prototype risk framework for risk specification, workflow simulation and analysis of preservation metadata is under development at the time of writing. This will be evaluated at the second DAVID test workshop in April 2015.

11 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. Available online: https://www.prestocentre.org/blog/8k-traffic-jam-ahead

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

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

[Avid, 2006] MXF Unwrapped, Avid Post Production, 2006. Available online: 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.

[Bailer, 2011] Werner Bailer, Hermann Fürntratt, Peter Schallauer, Georg Thallinger and Werner Haas, “A C++ Library for Handling MPEG-7 Descriptions,” in Proceedings of ACM Multimedia Open Source Software Competition, Scottsdale, AZ, USA, Nov. 2011, pp. 731-734.

[Bailer, 2014] Walter Allasia, Werner Bailer, Sergiu Gordea and Wo Chang, “A Novel Metadata Standard for Multimedia Preservation,” in Proceedings of iPres, Melbourne, AU, Oct. 2014.

[Bauer 2013] C. Bauer, JH.Chenot, H.Fassold (JRS), J.Houpert (CTI), Report on Usability-, Tools- & System- requirements, DAVID Deliverable D2.3, 2013.

[Bauer, 2014] C. Bauer. Initial Report on Validation and Tests. Deliverable D5.2, EC FP7 DAVID Project, 2014.

[Becker, 2010] C. Becker, H. Kulovits, and A. Rauber,Trustworthy Preservation Planning with Plato, ERCIM News 80, p.p. 24–25, Jan. 2010. Available online: 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. Available online: 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. Available online: 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. Available online: 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. Available online: http://hyperallergic.com/60598/eyebeam-hurricane-sandy-flooding/

[Chenot, 2013] Jean-Hugues Chenot and Christoph Bauer. Data damage and its consequences on usability, Deliverable D2.1, EC FP7 DAVID Project, 2013. Available online: http://david-preservation.eu/wp-content/uploads/2013/10/DAVID-D2-1-INA-WP2-DamageAssessment_v1-20.pdf

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

[Chivers, 2012b] L. Chivers, Shifted Crop Corruption, OPF Knowledge Base Wiki. Available online: 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. Available online: 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. Available online: 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. Available online: 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. Available online: 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. Available online: http://www.digitalpreservation.gov/formats/fdd/fdd000213.shtml

[DSpace] DSpace website. Available online: 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. Available online: 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. Available online: http://www.ninch.org/guide.pdf

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

[Hall-May, 2013] M. Hall-May, G. Veres, J-H. Chenot, C. Bauer and W. Bailer, DAVID D3.1 Initial Strategies and Risk Framework. Deliverable D3.1, EC FP7 DAVID Project, 2013. Available online: http://david-preservation.eu/wp-content/uploads/2013/12/DAVID-D3.1-ITInnov-Initial-Strategies-and-Risk-Framework.pdf

[Hall-May, 2014] M. Hall-May, B. Arbab-Zavar, J. Houpert, C. Tiensch, H. Fassold, and V. Engen.. Analysis of loss modes in preservation systems. Deliverable D2.2, EC FP7 DAVID Project, 2013.

[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. Available online: 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. Available online: 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 Solutions, 2010. Available online: 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. Available online: 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. Available online: 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. Available online: 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. Available online: http://blogs.loc.gov/digitalpreservation/2013/01/is-jpeg-2000-a-preservation-risk/

[Lloyd 2011] V. Lloyd *et al*. ITIL Continual Service Improvement – 2011 Edition. The Stationery Office, 2011. ISBN: 9780113313082.

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

[LOCKSS] Lots Of Copies Keeps Stuff Safe (LOCKSS) website. Available online: 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.

[Matlab] Matlab – The Language of Technical Computing. Available online: http://www.mathworks.com/products/matlab

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

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

[MPEG-7] ISO/IEC 15938-5, Information technology – Multimedia content description interface (MPEG-7) – Part 5: Multimedia description schemes

[MXF Legalizer] MXF Leglizer, Cube-Tec International. Available online:

http://www.cube-tec.com/products/mxf-legalizer/mxf-legalizer-text

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

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

[OWLSKOS] Using OWL and SKOS, Sean Bechhofer and Alistair Miles, May 2008, http://www.w3.org/2006/07/SWD/SKOS/skos-and-owl/master.html

[Panzer-Steindel, 2007] B. Panzer-Steindel, Data integrity CERN/IT, Draft 1.3, Apr. 2007. Available online: 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. Available online: http://www.loc.gov/standards/premis/

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

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

[PROV-O] PROV-O: The PROV Ontology, W3C Recommendation 30 April 2013. Available online: http://www.w3.org/TR/prov-o/

[RDF] RDF 1.1 Concepts and Abstract Syntax, W3C Recommendation 25 February 2014. Available online: http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/

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

[REWIND] EC FP7 268478 REWIND, Reverse Engineering of Audio-visual Content Data. Available online: 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. Available online: http://aisel.aisnet.org/acis2005/50/

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

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

[Rosenthal, 2009b] D. Rosenthal, Are format specifications important for preservation? Jan. 2009. Available online: 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. Available online: 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. Available online: 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] Signavio website. Available online: 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. Available online: http://www.nedcc.org/assets/media/documents/dman.pdf

[SKOS] SKOS Simple Knowledge Organization System, Reference, W3C Recommendation 18 August 2009, http://www.w3.org/TR/2009/REC-skos-reference-20090818/

[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. Available online: 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. Available online: 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. Available online: http://bpmcenter.org/wp-content/uploads/reports/2012/BPM-12-13.pdf

[Surridge 2012] M. Surridge, A. Chakravarthy, M. Hall-May, C. Xiaoyu, B. Nasser and R. Nossal. SERSCIS: Semantic Modelling of Dynamic, Multi-Stakeholder Systems. In, 2nd SESAR Innovations Days, 2012

[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. Available online: 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. Available online: http://forum.videohelp.com/threads/284080-Video-corruption-when-trying-to-convert-MPEG-2-to-DVD-compliant

[VidiCert] VidiCert, Joanneum Research. Available online: http://www.joanneum.at/en/digital/productssolutions/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. Available online: 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). Available online: 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. Available online: 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. Available online: 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. Available online: 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. Available online: http://wiki.opf-labs.org/display/REQ/Digital+Preservation+and+Data+Curation+Requirements+and+Solutions

[XCL] eXtensible Characterisation Language. Available online: 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. Available online: http://www.orc.soton.ac.uk/fileadmin/downloads/5D_Data_Storage_by_Ultrafast_Laser_Nanostructuring_in_Glass.pdf

12 Glossary

Terms used within this deliverable, sorted alphabetically.

Term	Description
Active Control	It is a procedure used when costs of Ad-Hoc control become very expensive and it includes global actions such as re-training staff or allocating more resources.
Ad-Hoc control	It is a procedure which allows to deal with any risk occurring on the fly.
AIP	Archive Information Package.
ARD	„Arbeitsgemeinschaft der öffentlich-rechtlichen Rundfunkanstalten der Bundesrepublik Deutschland“ (Association of Public Broadcasters in Germany)
AV	Audio-Visual.
AXF	Archive eXchange Format.
BDI	Belief, Desire, and Intention.
BPM	Business Process Model.
BPMN	Business Process Modelling Notation.
BPRisk	A short term for the Business Process Risk Management Framework.
COI	Cost Of Inaction.
Control	See ‘Risk Control’ below.
CRC	Cyclic Redundancy Check.
CRUD	Create Read Update Delete, a term commonly used for database functionality.
CSV	Comma Separated Values, a common file format.
CVaR	Conditional Value at Risk.
DAVID	Digital AV Media Damage Prevention and Repair.
Deming cycle	A cycle that represents a continual improvement process, also referred to as the PDCA cycle (Plan, Do, Check, Act).
DiMi	Digital Migration.
DIP	Dissemination Information Package.
Expected Success	A risk simulation term, representing a percentage that shows how successful Ad-Hoc control is for the one attempt (to repair/fix).
FFV1	FFmpeg video codec 1.
GOP	Group Of Pictures.
HD	High Definition.
HDD	Hard Disk Drive.
HSM	Hierarchical Storage Management.
HTML	Hyper Text Markup Language.
IMX	Sony Betacam IMX video format, using MPEG2 4:2:2 intra-frame compression at 30, 40, or 50Mbps. Available as cassettes, but can be exported as MXF files (known as MXF-D10).
IT	Information Technology.
ITIL	Information Technology Infrastructure Library, widely used and accredited best practices for IT service management.
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.

Term	Description
MP-AF	MPEG Multimedia Preservation Application Format.
MPEG	Moving Picture Expert Group.
Mitigate	To mitigate risk means to reduce the exposure to risk, to reduce its likelihood of occurring or its impact should it occur.
MoR	Management of Risk.
MTTF	Mean Time To Failure.
MXF	Material eXchange Format.
Negative Consequence	A risk simulation term, which is a quantitative measure of something going wrong during preservation process. It can be measured in percentages, monetary value or level of severity scale.
OAIS	Open Archival Information System.
OGC	Office of Government Commerce.
OP	Operational Profile.
Overhead	This is a cost paid for out some centralised fund.
OWL	The Web Ontology Language.
PDCA	Plan, Do, Check, Act.
PREMIS	Preservation Metadata: Implementation Strategies.
PROV	PROV is the term given to a family of documents that is part of the W3C Working Group on ‘provenance’, including, *inter alia*, a data model (PROV-DM), ontology (PROV-O), XML schema (PROV-XML) and a notation (PROV-N).
Provenance	A term generally referring to the origin or earliest known history of something, previously most commonly used in the context of art or antiques. However, it is also more widely used as a term that refers to historical information about a something, referred to as an ‘entity’, and all the activities and actors (e.g., people) involved in creating, modifying or destroying it. For example, the W3C PROV Working Group also refer to the provenance of data in computer systems.
QA	Quality Assurance.
QC	Quality Check/Control.
RAID	Redundant Array of Independent Disks.
RDF	Resource Description Framework.
REST	REpresentational State Transfer, an architectural design for web services.
Risk	Some process which can take place during normal workflow operation and lead to some unexpected changes to asset (digital item) or disruption of the operation.
Risk Control	Procedure which prevents consequences of risk occurred to proceed further in the workflow.
ROI	Return On Investment.
RTD	Research and Technology Development.
SD	Standard Definition.
SHA-1	Secure Hash Algorithm 1.
SIP	Submission Information Package.
SKOS	Simple Knowledge Organisation Scheme.
SMPTE	Society of Motion Picture & Television Engineers.
SOA	Service Orientated Architectures.
SPARQL	SPARQL Protocol And RDF Query Language.
SPIN	SPARQL Inferencing Notation.

Term	Description
SPOT	Simple Property-Oriented Threat (model), describing digital preservation properties which can be affected if a risk took place and no control actions taken.
SSD	Solid State Drive.
URI	Uniform Resource Identifier.
VaR	Value at Risk.
W3C	World Wide Web Consortium.
Workflow	The sequence of actions which is performed to achieve some goal in digital preservation process from initiation to completion.
Workflow task	It is a type of workflow action that determines the details of an assignment given to a specified worker(s) by a workflow rule or approval process. Also interchangeably referred to as a workflow ‘activity’.
XML	Extensible Markup Language.

Partner Acronyms

Term	Description
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

13 Annex A: BPMN 2.0 Overview

The purpose of this annex is not to describe BPMN 2.0 in detail, but to give a brief overview for readers who are not familiar with this standard modelling notation. Firstly, a generic workflow consists of nodes connected by edges like a graph. Figure 16 on page 61 is an example of a workflow, the MXF Repair workflow from ORF. An explanation of *nodes* of the archive model derived from BPMN2.0 model is provided below.

13.1 BPMN 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, which 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.

13.2 BPMN 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*.

14 Annex B: ORF MXF Repair specification details

The purpose of this Annex is to specify data which is essential for risk modelling using Risk simulation tool and two tasks from ORF MXF Repair workflow are given as an example. For a description of what each parameter means, please refer to Section 8.4 on page 40.

Parameter	TSM Retrieve		Mapping Control	
	Wrong File selected	Retrieve Fails	Overload	Wrong Assessment
Estimated Frequency	5 per year	5 per year	2 per month	1 per month
Estimated Frequency Factor	0.0004166	0.0004166	0.002	0.001
Multiplication Factor	1	1	1	5
Multiple Risk Entry per Task	No	No	Yes	Yes
Frequency of combined	None	None	None	None
Detection Level	0.9	1	1	0.75
Level of Severity	1	1	2	2
Ad-hoc effort available via Overhead	Yes	Yes	No	Yes
Expected Success of ad-hoc counter-measures	100%	90%	50%	90%
Cost associated with Risk, CAR per hour	€50	€50	€70	€50
Time spent on dealing with risk, TAR per item	0.1hrs	0.2hrs	0.5hrs	0.2hrs
Cost for active control strategy, CACS	€800 euros	None	€1000	€800
Active Control Activation rule	(CAR*TAR)*1.2 > CACS	None	(CAR*TAR)*1.0 > CACS	(CAR*TAR)*1.2 > CACS
Delay of effect, days	5	0	5	22
Availability	1	1	1	1
Identity	0	0	0	0
Persistence	0	0	0	0
Renderability	0	0	0	0
Understandability	0	0	0	0
Authenticity	1	0	1	1
Other	0	0	0	0

15 Annex C: Controlled vocabularies for activities, tools and risk

The following controlled vocabularies for activities and tools have been defined for qualifying the corresponding entities of the preservation metadata model. The following tables provide an overview of the defined terms. Serialised representations of these controlled vocabularies using MPEG-7 Classification Schemes and Simple Knowledge Organisation Scheme (SKOS) have also been created.

15.1 Specific activities

The following hierarchy of activities is derived from DAVID deliverable D2.3 [Bauer 2013] process descriptions, iModel and PrestoPRIME documentation.

Activity	Description	Specific properties
Preservation	Root activity for container activity representing an entire preservation workflow.	
Acquisition/Recording	The process of recording a signal from an external source to a media carrier.	
Ingest	The process of capturing, transferring, or otherwise importing different types of media content into a system.	
Storage	Storing ingested media data in the system.	
Migration/Distribution copy generation/proxy generation	Generating an instance of a media item as a copy of an existing media item, possibly in a different format or on a different media carrier.	
Analogue transfer recording	Creating a copy of an analogue media carrier on another analogue media carrier.	
Digitisation	Creating a digital copy of an analogue media carrier.	
Scanning	Creating a digital copy of film.	
Transfer Recording	Recording and digitising a signal from an analogue media carrier (e.g., magnetic tape)	
Digital Migration	Creating a copy of a digital media item in a different format or container.	
Wrapping	Creating a container holding one or more media streams.	
Transwrapping	Creating a container holding all media streams extracted from another container.	
Transcoding	Creating a copy of a media item in another format and/or encoding.	
Cleaning	Cleaning a media carrier.	
Checking	Checking the fixity, integrity or quality of a media item.	
Checksum	Checking fixity by using checksums.	

Activity	Description	Specific properties
Generation	Creating checksums from the media item.	
Verification	Verifying a media item against an existing checksum.	
Integrity	Checking that a media item is complete and unaltered.	
Quality control	Controlling the quality of the media item, its container and metadata.	
Human	Controlling visual and audio quality by displaying the content to a human operator (“eyeball check”).	
File-based	Automatic file-based quality analysis.	
Verification	Verification of automatic quality analysis results by an operator.	
technical	Verification of automatic quality analysis results by an operator w.r.t. technical aspects.	
editorial	Verification of automatic quality analysis results by an operator w.r.t. editorial aspects.	
Material selection	Selecting material based on defined criteria.	
Editorial	Editorial selection of material.	
Technical	Selection of material based on technical properties.	
Carrier, Format	Selection of material based on the type of carrier or file format.	
Condition	Selection of material based on its technical condition (e.g., based on quality analysis results).	
Material handling	Operations around physical movement of media carriers.	
Ordering	Placing an order for a media item.	
Check-out	Checking out an item in an archive management system.	
Taking	Fetching media carriers from shelves.	
Check-in	Checking in an item in an archive management system.	
Shelving	Putting media carriers on shelves.	
Shipping	Sending media carriers to their destination.	
Receiving	Media carriers arriving at their destination.	
Mapping control	Check the match between file content and metadata	
Preview control	Check quality and alignment of proxies generated for preview	
Carrier liquidation	Deleting content from media carriers.	
Revision	Mark media carriers that have already been migrated	

Activity			Description	Specific properties
	Deletion notification		Registering the deletion of a media carrier in an archive management system.	
Metadata			Operations concerning the metadata of media items.	
	Extraction		Extract metadata from audiovisual content.	
		Technical	Extract technical metadata from containers bitstream or essence.	
		Descriptive	Extract descriptive metadata from audiovisual essence.	
	Enrichment		Adding/completing metadata of a media item.	
		Technical	Adding/completing technical metadata of a media item.	
		Editorial	Adding/completing editorial metadata of a media item.	
	Modification		Modifying/correcting/updating existing metadata.	
	Conversion		Concert metadata from one representation into another one.	
Repair, Correction			Fixing problems of a media item by replacing it partially or entirely.	
	Replace copy		Replace a corrupted media item with an intact copy.	
	Local repair		Replace parts of a corrupted media item with an intact copy.	
	Re-encoding		Repeat encoding on a media item in order to fix encoding problems.	
	Rewrapping		Repeat wrapping on a media item in order to fix container problems.	
	Restoration		Apply corrections and improvements to restore an earlier condition of the media item.	
Access			Access a digital media item.	
	Delivery		Make available and transfer a digital media item.	
	Receive		Receiving a digital media item.	
Communication			Communication between agents in a preservation process.	
	Notification		Notify another agent about the status of a process or a media item.	
	Clearing		Reach agreement between agents about proceeding in the process.	
Auxiliary			Support activities to preservation processes.	
	File copy		Copy files between storages.	
	Temporary storage		Putting data on temporary storage	

Activity	Description	Specific properties
Cache clearing	Clearing temporary storage of a system.	
Network transfer	Transfer files over the network.	
Human intervention	Involving a human operator to assess the situation and decide about further steps.	
Packaging	Creating a package to be deposited or transferred.	
SIP creation	Create a package for submission to a preservation system.	
AIP creation	Create an archival package in a preservation system.	
DIP creation	Create a package for distribution.	
Emulation	Emulate an environment in order to reproduce legacy content.	
Custody	Activities related to the custody of a digital item.	
Takeover	Taking over custody of an item from another agent.	
Transfer	Transferring custody of an item to another agent.	

15.2Specific tools

Tool	Properties
Acquisition/Recording	EBU Tech 3349 (EBU Acquisition Technical Metadata Set) provides metadata for cameras. Many of the properties apply also to other acquisition devices.
Camera	
Film	
Analogue	
Digital	
Tape Recorder	
Analogue	
Digital	
Optical disk recorder	
Digital capture board	
Film scanner	
Telecine	
Checksum generator/verifier	• Type of checksum • Granularity ○ Byte offset and number of bytes

Tool			Properties
			o Start time and end time o Stream/track
Quality Control			EBU QC checks and parameters (profile as e.g. used by FIMS QA)
Viewer			• Type of viewing device • Spatial resolution • Colour space and depth
Coding, Wrapping			
	Encoder		
	Wrapper		
	Decoder		
	Unwrapper		
	Multiplexer		
	Demultiplexer		
Repair / Restoration			
Metadata Extraction			
	Technical		
	Descriptive		
		Automatic speech recognition	
		Optical character recognition	
		Machine translation	
File transfer			
Media Asset Management			
	Storage		
	Cataloguing		
	Search & Retrieval		

15.3 Risk vocabulary

Below is a risk vocabulary for preservation activities, which has been compiled by collaborative effort by IT Innovation, ORF and JRS in the DAVID project.

Activity	Risk	Control
Acquisition/Recording	Wrong acquisition format	Check and Redo / Transcode

Activity	Risk	Control
	Inadequate content quality	Check and Redo / Post-production
	Wrong metadata	Check and correction
	Missing metadata	Check and correction
Ingest	Incorrect content ingested	Checking
Storage	Capacity overload	Partly re-training
Storage	Faulty storage process	Check and Redo
Storage	Wrong file-name	Partly re-training
Storage	Wrong formats/codecs	Check and Redo / Transcode
Storage	Faulty formats/codecs	Check and Redo / Transcode
Migration/Distribution copy generation/proxy generation	Copies do not match	Checking
	Migration fails	Rerun
Analogue transfer recording	Clogging	Cleaning and/or baking of source
	Material is damaged	Check transfer machine, check source material / retraining
	Material is destroyed	Check transfer machine, check source material / retraining
Digitisation	Quality loss	Reselect format / quality parameters, re-digitise
Scanning	Material is damaged	Check scanner, check source material / retraining / Raise service rate
	Material is destroyed	Check scanner, check source material / retraining / Raise service rate
Transfer Recording	Material is damaged	Raise service rate / Re-Training
	Material is destroyed	Raise service rate / Re-Training
	Wrong content separation	Allocate more resources / Re-training
	Wrong IN/OUT	Allocate more resources / Re-training
Digital Migration	Quality loss	Reselect format / quality parameters, re-migrate
	Information (metadata) loss	Switch migration tool, reintroduce missing metadata
	Incorrect information (metadata) addition	Correction
Transwrapping	Metadata taken from wrong source field	Correction, change trans-wrapping tool

Activity			Risk	Control
			Metadata added to wrong target field	Correction, change trans-wrapping tool
			Metadata not transferred	Correction, change trans-wrapping tool
			Stream written to wrong part of container	Correction, change trans-wrapping tool
		Transcoding	Field order issues	Correction, change transcoding tool
			Stream order problem	Correction, change transcoding tool
Cleaning			Material is damaged	Higher service rate / Re-Training, other cleaning method
			Material is destroyed	Higher service rate / Re-Training, other cleaning method
Checking			Wrong method	Other method
	Checksum generation		Wrong checksum generated	Generate redundant checksums
	Checksum verification		Check against wrong checksum	Retry, manual investigation
	Integrity checking		Not standards compliant	Rewrapping / re-transcoding (if small number of files), correction (if many), retraining (depending on source of file)
Checking			Wrong method	Other method
	Checksum generation		Wrong checksum type generated	Find a single common checksum type (e.g. MD5, 128bit)
	Checksum verification		Check against wrong checksum	Retry, manual investigation Correct checking workflow
	Integrity checking		Not standards compliant	Rewrapping / re-transcoding (if small number of files), correction (if many), retraining (depending on source of file)
	Quality Control			
		Eyeball-check-based	Missing important errors	Increase observation manpower, add semi- automatic video-based tests
		File-based	Wrong settings	Manual QC / Re-Tests / Re-Grouping
			Overload	Increase engine performance
			QC works inaccurate	Engage QC specialist, adapt or change QC engine
			Higher complexity than expected	Engage QC specialist, adapt or change QC engine
			New file type (codec or wrapper)	Engage specialist, adapt or change QC engine
		Verification		

Activity	Risk	Control
Technical	Wrong parameter	Manual QC / Re-Tests / Re-Grouping
	Overload	Increase engine performance
	QC works inaccurate	Engage QC specialist, adapt or change QC engine
	Higher complexity than expected	Engage QC specialist, adapt or change QC engine
Editorial	Faulty assessment	Re-Adjustment VidiCert / Re-Grouping / Re-Training
	Too much information	Redesign QC workflow Engage QC specialist, adapt or change QC engine
	Process more complex than expected	Engage QC specialist, adapt or change QC engine
Material selection	Wrong content selected	Re-Training / Redo Criteria-list / Re-selection / Re-Grouping
	Wrong source-material selected	Re-Training / Redo Criteria-list / Re-selection / Re-Grouping
	Relevant content not selected	Re-Training / Redo Criteria-list / Re-selection / Re-Grouping
	Too little amount of output	Allocate more man-power
Material handling	N/A	
Ordering	Orderlists are not submitted	Logistic changes / Changes in info-flow /
	Too many items in Orderlist	Re-Training / Changes in info-flow
	Too few items in OL	Re-Training / Changes in info-flow
	Non-ideal composition of OL	Re-Training / Changes in info-flow
	Wrong selection timing	Re-Training / Changes in info-flow
Taking	Wrong material is taken	Re-Training / Logistic changes / Changes in info-flow / Search-party
	No empty carts available	Purchase more carts
	Material is missing	(no controls available)
Check-out	Fails	Re-Do
Shelving	Wrong tapes in shells	Re-Training
	Faulty shelving	Re-Training
Check-in	Fails	Re-Do

Activity	Risk	Control
Shipping	Climatic shock	Logistic changes / provide conditioned shipping-units
	Accident small	Re-Training / Logistic changes
	Accident big	(no controls available)
	Material gets temporarily lost	Re-Training / Logistic changes
	Logistic Failures	Re-Training / Logistic changes
Mapping control	Too many errors to be corrected	Re-Training / workflow-changes
	Faulty assessment	Re-Training
	Overflow	Allocate more man-power
Preview control	Wrong assessment	Re-training
	Overflow	Allocate more man-power
Carrier liquidation	Wrong material is deleted	Re-training
Deletion notification	Wrong notification	Re-training
Metadata	Overflow	Allocate more man-power
Enrichment	Wrong enrichment	Re-training
Modification	Wrong modifications	Re-training
Repair, Correction	Fails	Other method
Replace copy	Not available	Other method
Local repair	Fails	Other method
	Takes too long	Move to restoration
Restoration	Fails	Other method / hand over to expert
Access		
Delivery	Climatic shock	Logistic changes
	Accident small	Re-Training / Logistic changes
	Accident big	(no controls available)
	Material gets temporarily lost	Re-Training / Logistic changes
	Logistic Failures	Re-Training / Logistic changes
	Fails	Other method of delivery
Receive	Receiving “station” fails	Re-Try
Communication		
Notification	Notification is not given	Logistic changes / Changes in info-flow
	Wrong notification	Re-Training

Activity		Risk	Control
	Clearing	Sent for wrong (faulty) material	Re-Training / changes in info-flow
		NOT sent	Logistic changes / Changes in info-flow
Auxiliary			
	Cache clearing	Wrong files are deleted	Re-Training / changes in info-flow

16 Annex D: Preservation Metadata Service API

This annex details the API for the preservation metadata service discussed above in Section 8.6 on page 52.

16.1 Definitions

A **workflow instance** is the execution of a specific workflow for one media item. Repeating the entire workflow for the same item will result in a new instance. The preservation metadata service only contains data about workflow instances that are no longer running, i.e., are completed, failed or were aborted.

A **preservation metadata document** contains metadata about one workflow instance or a group of workflow instances. It describes the preservation actions applied to a media item as well as their parameters.

16.2 Metadata Representation

The metadata representation is the XML format defined by MPEG Multimedia Preservation Application Format (MP-AF). In addition, the following restrictions apply:

- Metadata documents describing multiple workflow instances must be modular, i.e., it must be possible to take the members of the Group element on the top level and use them as independent metadata document. This means, that no references between workflow instances may be used.

Identification of workflow instances: Each description of a workflow item must contain a ProvenanceDescriptor with a top-level Activity representing the entire workflow. The `uri` attribute of the Activity shall correspond to “urn:uuid:” followed by a UUID for the workflow instance. If the workflow manager does not provide unique identifiers, the preservation metadata service shall coin a URI that prefixes the workflow instance ID in order to make it unique. The `type` attribute of the Activity shall identify the workflow definition, either as UUID or using the URL of the BPMN file defining the workflow.

16.3 Web service functionality

The service provides a RESTful web service interface with the following functionality.

GET /service/version

Description: Returns the current version string of the service in the form "major.minor.patch".

Parameters: none

Return:

- Content-Type: text/plain
- Content: a plain text line with the version string.

POST /pmdoc

Description: Adds the preservation metadata document specified in the content. If the preservation metadata document is already in the system, it will be replaced with the new preservation metadata document.

Parameters: none

Accepts:

- Content-Type: application/xml

Return:

- Content-Type: text/plain
- Content: the id of the preservation metadata document (the id of the workflow instance, or a unique id derived from the workflow instance id)

GET /pmdoc

Description: Returns a preservation metadata document with a description of the workflow instances matching the query.

Parameters:

- optional *timeStartAfter*: if specified, all workflow instances started after the specified date (in ISO 8601 Format) are returned.
- optional *timeEndBefore*: if specified, all workflow instances completed before the specified date (in ISO 8601 Format) are returned.
- optional *timeAddedAfter*: if specified, all metadata documents added to the service after the specified date (in ISO 8601 Format) are returned.
- optional *timeAddedBefore*: if specified, all metadata documents added to the service before the specified date (in ISO 8601 Format) are returned.
- optional *systemId*: if specified, all metadata executed on the system with the specified id are returned.
- optional *workflowId*: if specified, only metadata documents containing instances of the specified workflow instance
- optional *idsOnly*: Return only the list of ids of the documents. Defaults to true.
- If more than one time parameter is specified, they will be combined using logical AND.
- If no *time**, *workflowId* or *systemId* parameters are specified, the metadata of all workflow instances in the system in the system will be returned.

Return:

- Content-Type: application/xml
- Content: If *idsOnly* is specified, the list of workflow instances surrounded by <id> and grouped in a <pmdoc> element, otherwise a document conforming to MPEG MP-AF, containing a group of preservation metadata records (each describing one workflow instance)

Examples:

- GET http://server/pmdoc?timeStartAfter=2014-11-07T13:00:00
- GET http://server/pmdoc?timeEndBefore=2014-11-07T13:00:00
- GET http://server/pmdoc

GET /pmdoc/{id}

Description: Returns the preservation metadata document for the workflow with the specified id.

Parameters: none

Return:

- Content-Type: application/xml
- Content: A document conforming to MPEG MP-AF, containing preservation metadata for one workflow instance, or HTTP status 404, no metadata for a workflow with this ID exists.

GET /workflowDefIds

Description: Returns the list of workflow definition ids, for which workflow instances are available.

Parameters: none

Return:

- Content-Type: application/xml
- Content: A list of workflow definition ids, each enclosed in a <id> tag, and the list enclosed in <ids>. The list may be empty.

DELETE /pmdoc/{id}

Description: Deletes the preservation metadata document for the specified workflow instance.

Parameters: none