Metadata Representation and Risk Management Framework for Preservation Processes in AV Archives

Werner Bailer
JOANNEUM RESEARCH – DIGITAL
Steyrergasse 17
8010 Graz, Austria
+43 316 876 1218
werner.bailer@joanneum.at

Martin Hall-May, Galina V. Veres
University of Southampton – IT Innovation Centre
Gamma House, Enterprise Road
SO16 7NS, Southampton, United Kingdom
+44 23 8059 8866
{mhm,gvv}@it-innovation.soton.ac.uk

ABSTRACT
This paper proposes an approach to assessing risks related to audiovisual (AV) preservation processes through gathering and representing metadata. We define a model for process metadata, which is interoperable with both business process models and other preservation metadata formats. A risk management framework is also suggested to help key decision makers to plan and execute preservation processes in a manner that reduces the risk of 'damage' to AV content. The framework uses a plan, do, check, act cycle to continuously improve the process based on risk measures and impact model. The process metadata serves as the interface between the steps in the framework and enables a unified approach to data gathering from the heterogeneous tools and devices used in an AV preservation workflow.

General Terms
Infrastructure, preservation strategies and workflows.

Keywords
Process metadata, business processes, risk management, risk assessment, simulation.

1. INTRODUCTION
Preservation processes for audiovisual content consist of complex workflows involving numerous interrelated activities performed by different tools and devices. Interoperable metadata throughout the entire workflow is a key prerequisite for performing, monitoring and analysing such preservation processes.

2. METADATA REPRESENTATION
For preservation purposes two types of metadata are most crucial: structural metadata (technical metadata needed to correctly interpret the stored essence) and preservation metadata (metadata for assessing the fixity, integrity, authenticity and quality of the object, as well as a documentation of the preservation actions applied). While the first is sufficiently covered by many existing formats, there is still a gap for representing preservation metadata for AV preservation processes. This paper focuses on the second type of metadata. Such processes as ingest, digitisation or migration can be quite complex, and heterogeneous workflows involve a number of different devices, software tools/ systems and users. We propose a metadata model for documenting the procedures applied to multimedia content in a preservation process together with tools, their parameters and operators involved. These metadata can be used for different applications, such as automatically adapting preservation and restoration workflows/tools, or collecting data for the assessment/simulation of risks related to these processes.

The scope of the preservation process metadata model is to document the history of creation and processing steps used, 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. It supports a set of specific types of activities in the model (e.g., digitisation) with possible further specialisations (e.g. film scanning) in order to improve interoperability between preservation systems. The model also describes the parameters of these activities. There is a core set of well-defined properties together with their types, which store the value used when processing the item described. In addition, a key/value structure for supporting extensions is provided.

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 *DigitalItem* represents an intellectual/editorial entity to be preserved. This entity has been borrowed from the MPEG-21 Digital Item Declaration (DID) model [1]. A DigitalItem aggregates other DigitalItems, such as the representations of an intellectual/editorial entity and the essences constituting the representation, and *Components*, such as the bitstreams of an essence. A *Component* is the binding of a resource to a set of metadata. It is not an item in itself, but a building block of items. It aggregates *Resources*, which are individually identifiable content files or streams in a container. A resource may also potentially be a physical object. All resources shall be locatable via an unambiguous address. Specialised subclasses of DigitalItem (such as supported in MPEG-21 and PREMIS [3]) can be optionally added, but are not needed for the purpose of describing preservation history. The model allows describing DigitalItems and Components without related Resources, which is useful for describing preservation activities that failed and left no trace in form of essence, but have to be documented for risk assessment.

An *Activity* is an action in the lifecycle of the content item which creates, uses or modifies a DigitalItem. Activities may be composed of other fine-grained Activities. 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. Thus we achieve a simpler representation than in process models such as BPMN [2]. 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.

An *Operator* is an entity contributing to the completion of an Activity by performing it or being used to perform it. The type of involvement is further specified by the Operator's role attribute. An Operator is either an *Agent* (a person or organisation involved in performing an activity) or a Tool (a device or software involved in performing an activity). The description of tools includes parameters and resource usage information. Operators may act on behalf of other Operators (e.g., Tools being used by Agents).

The metadata model constitutes a subset of the MPEG MP-AF data model described in [5].

3. RISK MANAGEMENT FRAMEWORK

We propose a risk management framework to help key decision makers to plan and execute preservation processes in a manner that reduces the risk of 'damage' to AV content. Damage is considered to be 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.

This risk management framework relies on a repetitive procedure of planning and simulation of preservation processes, adapting and executing them, and gathering data from the execution for updating the risk model and simulation. Data gathering requires breaking down the process model, which contains all possible execution paths, into the sequence of actions that have actually been executed. Then the data are collected from configuration and execution logs of the individual tools. These data have to include not only operational information, but also risks-related knowledge. This knowledge consists of identified risks and their frequency of occurrence, their negative consequences and effects on assets (AV content), any controls dealing with the risks and their associated time and cost.

A cycle of continuous process improvement is proposed, which involves the following steps: plan, do, check, act. The basis of planning decisions is a simulated business process, representing the critical activities, tools and properties of the key preservation workflows (ingest, migration and access). The critical part to such a risk management approach is to ensure that the models and simulations of business processes used for planning decisions are kept consistent with the actual execution.

Most tools available for business process modelling are generic, offering no particular guide to the modeller. We use a controlled vocabulary to help to design the workflow, describe risks and thereby synchronise with the execution model. It also allows us to relate data gathered from the executing process to the activity in the workflow and to determine when and how risk measures are being breached. Three risk measures are suggested for preservation processes: Expected loss (mean of negative consequences (NC) which can occur in a given process), Value at Risk (minimum NC incurred in α% of the worst cases in a given process) and Conditional Value at Risk (mean of NC incurred in α% of the worst cases). These risks measures can be calculated allowing both propagation of risks through the preservation process and usage of controls to deal with risks occurred.

The metadata model is the interface between simulation and execution, as it allows us to map from abstract preservation activities, tools and their significant properties to and from their actual implementation. Metadata on process execution can be gathered for statistical analysis, and allows us to monitor preservation workflows in a manner that is consistent with planning models.

The purpose of the risk management framework is to allow the archive decision-makers to balance the cost and time involved in avoiding and mitigating risks with the risk reduction achieved by deploying 'controls' in the business process. By closing the loop between simulation and execution, the reliability and accuracy of the data used to drive planning decisions is improved, which is critical to justify any additional expenditure for uncertain future gains (i.e. long-term access to content).

To classify the impact of risks in digital preservation, we use the Simple Property-Oriented Threat Model (SPOT) as an impact model for Risk Assessment. The SPOT model [4] defines six essential properties of digital preservation: Availability, Identity, Persistence, Renderability, Understandability, and Authenticity.

The implemented demonstrator uses the metadata model to represent the data gathered from process definitions and execution logs and runs simulations using the risk assessment.

4. CONCLUSION

The proposed approach enables decision makers in AV preservation to make their decisions based on information about the risks involved. The risks can be assessed and simulated not only on estimates but on actual data gathered from the execution of preservation processes. This will provide a much more realistic and reliable assessment of risks and thus allow the risks of a damage to audiovisual content to be better managed.

5. ACKNOWLEDGMENTS

This work has been funded partially under the 7th Framework Programme of the European Union within the ICT project "DAVID" (ICT FP7 600827).

6. REFERENCES

[1] ISO/IEC 21000-2, Information technology – Multimedia framework (MPEG-21) – Part 2: Digital Item Declaration

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

[3] PREMIS Editorial Committee, 2008. *PREMIS Data Dictionary for Preservation Metadata*, version 2.0, http://www.loc.gov/standards/premis/v2/premis-2-0.pdf

[4] Vermaaten, S., Lavoie, B., and Caplan, P. 2012. Identifying Threats to Successful Digital Preservation: the SPOT Model for Risk Assessment, *D-Lib Magazine*. 18, 9/10, 2012.

[5] Allasia, W., Bailer, W., Gordea, S. and Chang, W. A Novel Metadata Standard for Multimedia Preservation, *Proceedings of iPres*, Oct. 2014.