DIGITAL - Institute for Information and Communication Technologies

Collecting preservation metadata for risk assessment

Werner Bailer

EBU MDN Workshop, 9 June 2015

Outline

- Motivation, use cases
- Metadata representation
- Risk assessment in preservation
- Prototype implementation
- Conclusion

Motivation

- Preservation processes for audiovisual content consist of complex workflows
- Activities are performed by different tools and devices
- Planning and improving workflows requires assessment of related risks
- Interoperable metadata is a key prerequisite for performing, monitoring and analysing such workflows

Use cases for process metadata

- Preservation decisions
  - Decide on restoration/improvement based migration history
  - Select encoding parameters based on process and material properties in previous generations
- Restoration: select/configure restoration tools based on
  - device/tool
  - material properties
  - information/measurements from preservation activities (e.g. known issues of certain tape machines, side effects of previous error concealment tools applied)

Use cases for process metadata

- Data gathering for risk assessment
  - for items with same/similar problems
    - which activities & operators were involved?
    - are there patterns?
  - get statistics of activities/tools
    - failure rates
    - resource consumption

Metadata representation

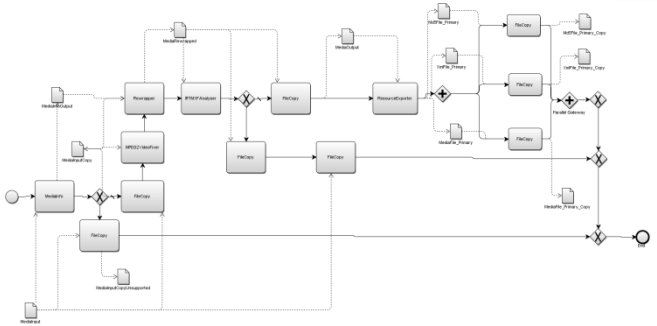
- Technical metadata
  - properties of the essence
- Preservation metadata
  - assessing fixity, integrity, authenticity and quality
  - documentation of the preservation actions applied

Metadata representation

- Tools in the workflow use a variety of different metadata formats
- Sources of metadata about process
  - Process models (planned workflow)
  - Log information (actually executed workflow)
  - Output metadata (results of checks, validation, etc.)

Process models

- Business process representations
  - widely used for preservation workflows
  - using standards like BPMN



- Level of modelling detail differs between organisations
  - subprocesses vs. tasks (disjoint types in BPMN)
  - often created as a conceptual model, lacking many details (e.g., specifying inputs/outputs of activities)

Process models

- Describes all possible branches of a workflow
  - including all error cases
  - may describe loops and recursions, which are flattened to a sequence in the execution
  - conditions, messages, etc. are not needed unless actually encountered during execution
- Lack details about actual execution
  - which implementation of a service has been called
  - which parameters/settings

Process models

- Preserving BPMN?
  - severe interoperability issues between different tools supporting BPMN
    - http://www.omgwiki.org/bpmn-miwg/doku.php
  - duality of the format
    - executable process vs. graphical representation
    - source of inconsistencies and problems

Log information

- Very heterogenous type of metadata
  - requires vendor specific implementation
- Option to obtain information
  - parse log files
  - custom APIs or web service interfaces
- What has been executed when, with which parameters?
  - time and machine of tool/service execution
  - version of tool/service
  - input parameters

Log information

- Workflow level
  - orchestration system
  - usually easier to obtain and better structured
  - often lacks detail (not reported back from indivisual service/tool)
- Tool/Service level
  - needed information is available
  - often very unstructured (textual log messages as they are output)
  - requires implementation for each tool, and maybe some reverse engineering

Output metadata

- Metadata produced by tools/services
  - metadata extractors
  - file validators
  - fixity and integrity checks
  - QC tools
  - signal parameters from tape machines
  - …

Output metadata

- Heterogeneous formats
- Some chance for standard formats in future
  - e.g., FIMS QA and AME reports
- Not everything needs to be processed in detail
  - success/failure (and reason) are most relevant

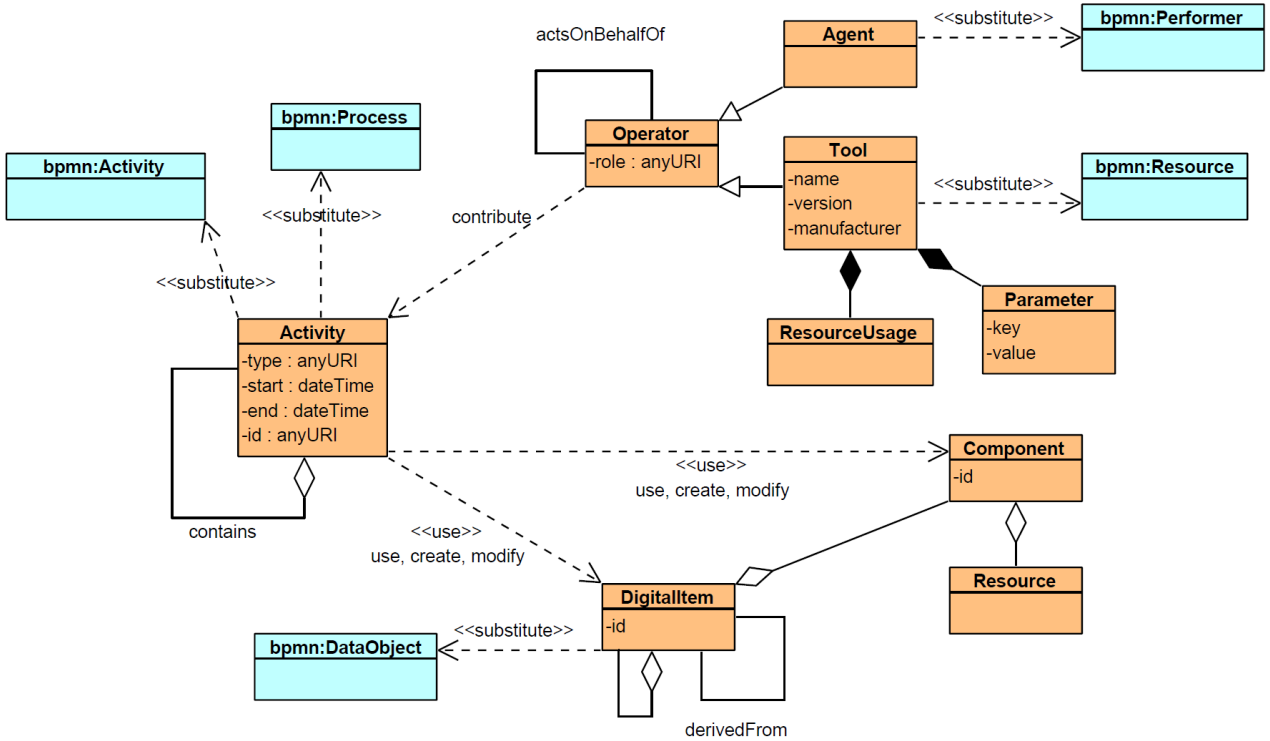
Metadata model

- Interface between various tools/services in the workflow and applications using metadata
- Ensure interoperability
  - avoid creating dependency on vendor specific format in subsequent applications
- Describe
  - activities
  - their relation to content
  - tools/services and agents involved
  - parameters/settings of tools/services

Metadata model

- Centered around
  - (Digital) Items
  - Activities
  - Operators
- Similar core structure as W3C Provenance model and PREMIS

Process metadata model



BPMN compatibility

- Process corresponds to an activity without parent activity
- Task corresponds to an activity without child activities
- Sub-process corresponds to an activity with child activities
- BPMN performer defined as Agent
  - linked to an Activity via a resource role
- BPMN resource is linked to an activity via a resource role
  - provides list of parameters indirectly via ParameterBinding

Compatibility with MPEG MP-AF

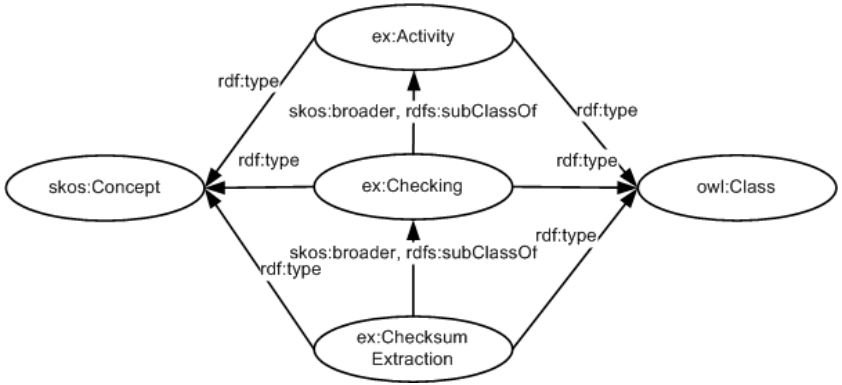
- Metadata model constitutes a subset of MP-AF
- MPEG MP-AF
  - application format for preservation
  - MPEG-21 DID container
  - includes eight groups of preservation metadata
  - using other MPEG standards (MPEG-7, MPEG-21), EBU Core and Dublin Core

Preservation metadata in MPEG MP-AF

- **Provenance**: creation, custody, (processing) history
- **Context**: circumstances of the production
- **Reference**: identifiers, linking related resources
- **Quality**: assessment of technical quality
- **Integrity**: presence, persistence of the resource
- **Authenticity**: assess authenticity of the resource
- **Fixity**: ensuring that properties have not been altered
- **Rights**: legal/contractual provisions affecting ownership, control, use

Data model implementation

- XML Schema
- Controlled vocabulary of Activity and Tool types
  - MPEG-7 classification scheme and SKOS representations
  - overlay approach to allow types to be SKOS concepts and OWL classes
    - least impact on either side
    - types of relations from the “other” model can be excluded to simplify reasoning (i.e., not require OWL Full)



Risk-aware business process management

- Static / design-time risk management
  - during design time (prior to execution)
- Run-time risk management
  - monitor the emergence of risks and apply mitigation actions during execution
- Off-line risk management
  - identify risks from logs and other post-execution artefacts

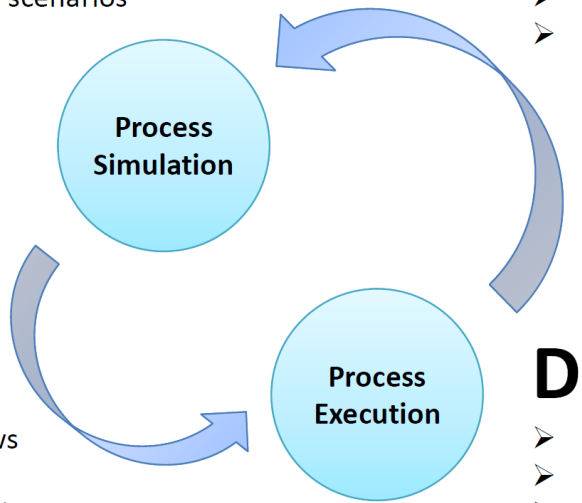
Risk management cycle

Plan

- Build workflows
- Capture risk information
- Simulate future scenarios
- Make decisions

Check

- Preservation metadata
- Process analytics
- Calibrate simulation
- Trigger live alerts



Act

- Adapt workflows
- Manage risks
- Configure risk alerts

Do

- Execute business processes
- Orchestrate services
- Record execution metadata

Steps in risk management cycle

- Identify workflows and processes taking place during digital preservation in archives
- Define objectives of risk management for digital preservation in archives
- Identify risks in workflows/processes, negative consequences and their effects on assets
- Identify controls dealing with risks and any associated costs and time

Risk classification (1)

- Simple Property-Oriented Threat (SPOT) model
- Impact model for risks
- Six essential properties of successful preservation

Risk classification (2)

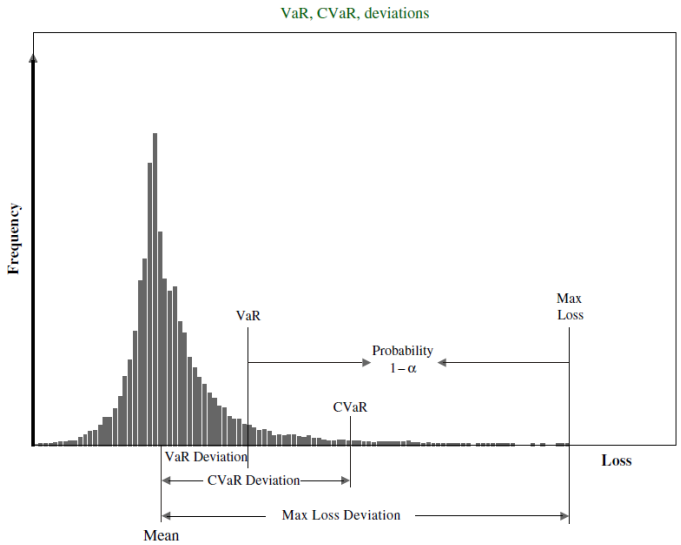
- Availability
  - digital object is available for long-term use
- Identity
  - digital object is referencable (can be distinguished from other objects)
- Persistence
  - the bit sequences continue to exist in usable/processable state and are retrievable/processable

Risk classification (3)

- Renderability
  - digital object is able to be used in a way that retains the object's significant characteristics, content, context, appearance, and behaviour
- Understandability
  - enough supplementary information such that the content can be appropriately interpreted and understood by its intended users
- Authenticity
  - digital object is what it purports to be

Risk measures

- Expected loss (E): average magnitude (mean) of negative consequences
- Value at Risk (VaR): minimum negative consequence incurred in α% of worst cases
- Conditional Value at Risk (CVaR): expected negative consequence incurred in α% of worst cases



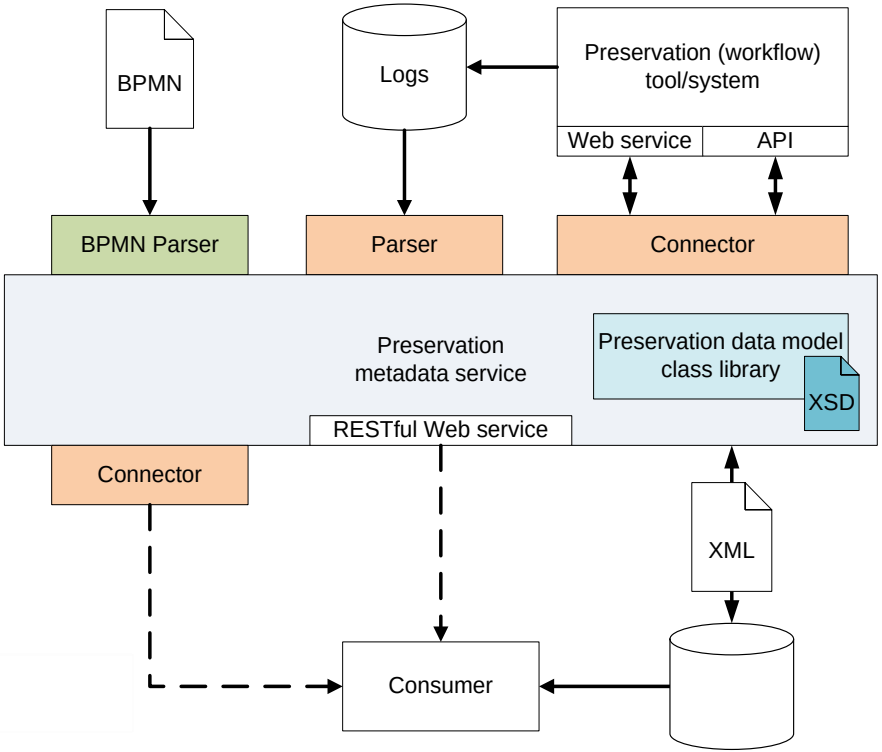
Data gathering

- Use metadata model as interoperable representation of information from different tools
- Gather data from configuration, workflow engines and logs
- Include data about
  - choices in workflow
  - exception handling
  - planned but not executed activities

Prototype implementation

- Tools
  - Library to handle data model
  - Gathering data from BPMN
  - Connector/parser components to get execution data (logs)
  - Access components
    - RESTful web service to provide data for risk assessment tools

Prototype implementation



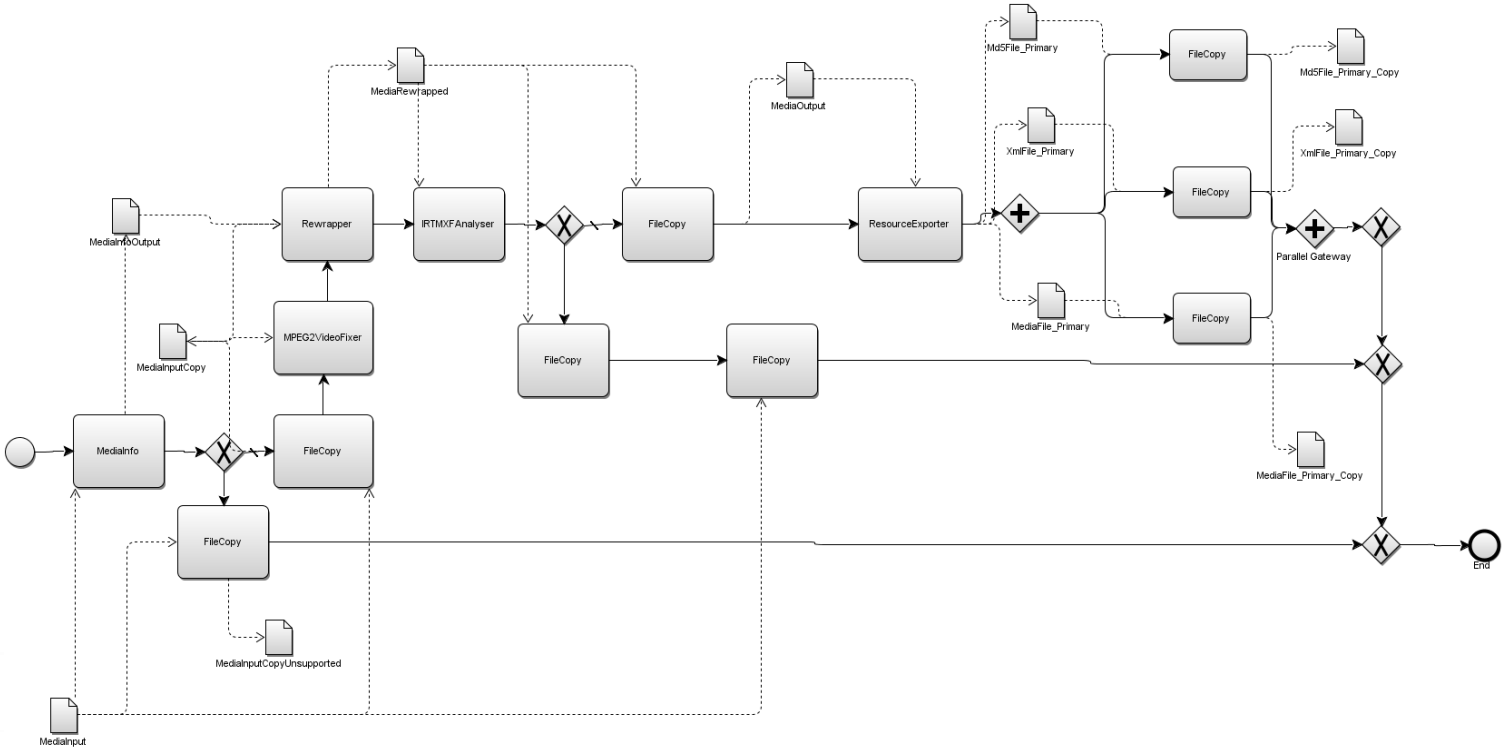
Prototype implementation

- Transform BPMN to MP-AF (1)
  - Execution flow
    - not directly modelled in MP-AF
    - model via default flows
      - set executed attribute to false
      - change when actually invoked
  - Data flow
    - include data objects in BPMN
    - essence vs. non essence: added during transformation
      - not in BPMN – specify metadata objects via parameters

Prototype implementation

- Transform BPMN to MP-AF (2)
  - Type classifications of activities
    - Activity type needs mapping to controlled vocabulary
  - Operators: not defined in BPMN
  - Implemented using XSLT
  - Result
    - MP-AF stub with possible activities
    - dummy IDs to be replaced with actual content identifiers

Example business process



Prototype implementation

- Handling metadata documents
  - C++ class library
  - generated from XML schema
- Support for entire MPEG MP-AF standard
  - to be published under LGPL very soon

Prototype implementation

- Implementing data gathering from log files
  - in our example, CubeWorkflow, MediaInfo, MXF Analyser
- Turn MP-AF stub into actual instance
  - content instances involved
  - tasks executed and their parameters
  - start/end times of tasks

Prototype implementation

- Workflow instance
  - execution of a specific workflow for one media item
  - repeating the entire workflow for an item results in a new instance
  - we only consider completed instances
- Preservation metadata document
  - contains metadata about one workflow instance or a group of workflow instances
  - describes the preservation actions applied to a media item as well as their parameters

Prototype implementation

- Service for preservation metadata provision
  - get metadata documents
  - different filter criteria (time, workflow)
  - get list of executed workflow types
- Results are provided as MPEG MP-AF compliant XML documents

Results (1)

- 1,135 items processed in workflow
  - regular workflow contains 10 activities
- 92 items did not go pass through the entire workflow
  - 37 activities failed, affecting 17 items
  - 3 items went through error handling path defined in workflow
- for 24 items one or more activities were executed more than once

Results (2)

- Activities failing most frequently
  - MediaInfo (7 items)
    - first activity in the process
  - Rewrapping (6 items)
    - appear for subsequent items
  - Different file copy steps (4 items)
    - seem to appear randomly

Results (3)

- Lessons learned for the process design
  - several failure cases not handled by exception paths in the workflow
  - often multiple activities fail in a sequence
    - could be avoided by exception path instead of executing dependent activities
  - some failed activities are rerun and produce results
    - not explicitly documented in the results

Conclusion

- Interoperable metadata format to gather heterogenous metadata from tools/services in preservation workflows
- Input for applications such as risk assessment
- Prototype implementation for a real-world migration workflow

Werner Bailer
JOANNEUM RESEARCH – DIGITAL

werner.bailer@joanneum.at
http://www.joanneum.at/en/digital

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreements n° 600845 (Presto4U) and n° 600827 (DAVID).