



Project Acronym: Presto4U
Grant Agreement no: 600845
**Project Title: European Technology for Digital
Audiovisual Media Preservation**

D3.3: Research Outputs Assessments v2

Project funded by the European Community in the 7th Framework Programme



Table of contents

Scope.....	4
Executive summary	5
1 Research Outputs Identification.....	6
1.1 Research Outputs Chosen	7
1.1.1 Metadata mapping.....	7
1.1.2 Vocabulary mapping.....	7
1.1.3 Quality assessment	8
1.1.4 Technical Metadata Extractors	10
1.1.5 Preservation and Platform Systems	10
2 Assessment Criteria Updates	12
2.1.1 Metadata mapping.....	12
2.1.2 Vocabulary mapping.....	18
2.1.3 Quality Assessment.....	23
2.1.4 Technical Metadata Extraction Template	33
2.1.5 Preservation Platforms Assessment Criteria	42
3 Results of Research Outputs Assessment – Year 2	52
3.1 Metadata mapping.....	52
3.1.1 Assessment results for Metadata Interoperability (MINT) toolset for EBUCore 52	
3.1.2 Assessment results for PrestoPRIME Metadata Mapping Tool	59
3.2 Vocabulary mapping.....	67
3.2.1 Assessment results for Amalgame	67
3.2.2 Test results.....	70
3.3 Quality assessment	70
3.3.1 Assessment results for VidiCert	70
3.3.2 Assessment results for BAVC QC Tools	74
3.3.3 Summary of Quality Assessment Tool Evaluation.....	78
3.4 Technical metadata extractors.....	78
3.4.1 Test environment.....	78
3.4.2 Dataset.....	79
3.4.3 Functional suitability	79
3.4.4 Performance efficiency	85
3.4.5 Resource Utilization	86
3.4.6 Compatibility (Interoperability)	89
3.4.7 Usability.....	89
3.5 Preservation Platforms Evaluation.....	92
3.5.1 DSpace.....	93
3.5.2 RODA.....	105
3.5.3 Archivematica.....	117
4 Final Presto4U Dataset.....	130
4.1 ICoSOLE	130
4.2 BLIP10000 Dataset	131
4.3 Dataset for MXF tool chain testing.....	132
5 Conclusion.....	134
Glossary.....	135
6 References	136
Annexes.....	137
Document information.....	138

Scope

The long-term preservation of digital audio-visual media presents a range of complex technological, organisational, economic and rights-related issues, which have been the subject of intensive research over the past fifteen years at national, European and international levels. Although good solutions are emerging, and there is a large body of expertise at a few specialist centres, it is very difficult for the great majority of media owners to gain access to advanced audio-visual preservation technologies. This deliverable 'Research Outputs Assessment v2' will describe the research outputs identified in year 2 of Presto4U, and which have the potential to address CoP needs and requirements. This document will also describe in detail, results of the assessment exercise carried out on identified ROs. The methodology for assessing these tools has been established as part of WP3 task T3.1 'Research Outcomes Assessment Methodology' documented in deliverable D3.1 'Specification of Assessment Criteria, Metrics, Processes, Datasets and Facilities'.

The deliverable is a direct outcome of Task 3.2 'Preservation Research Technology and Assessment'. The purpose of which is to identify and assess research outputs to establish their readiness for take-up. This deliverable is an update on D3.2 -- the lessons learnt in terms of assessment of tools from Year 1 are presented here. This is through updated test templates and improved tests in terms of their functional testing. Some of the tools needed to be re-tested in year 2 because of the updated test assessment templates. We also introduce two new categories of tools technical metadata extractors and vocabulary mapping. Further, the final Presto4U dataset is also described as part of this deliverable. A combination of open source free to download datasets and in-house produced files turned out to be the ideal dataset for the testing of these ROs. Finally we present the results of the tool testing performed after collaboration with commercial vendors to test two hardware storage mechanisms (LTO6 and Optical Drives).

Executive summary

The long-term preservation of AV media presents several challenges in terms of research and development required, rights related issues and the methodology required to assess the tools based on new and existing standards. The issue related to assessment of AV preservation tools is particularly challenging because within Presto4U the assessment methodology must also take into consideration the needs expressed by the CoPs. As the goal here (goal of the tools being produced by solution providers, software vendors) is for long-term preservation and given the fact that technology cycles are relatively short, there is a need for tracking and mapping of candidate solutions to keep up with on-going information technology developments. In Year 1 of Presto4U, we focussed on defining a methodology for the assessment of Research Outputs (ROs) (D3.1 [1]) and completed the first round of testing based on the templates defined for assessment as part of the methodology, the results of which were presented in D3.2 [4] .

In this deliverable, we present the results of the second round of RO assessment. The templates for assessment have been updated in terms of the criteria for assessment and their underlying functions (e.g. functional completeness, operability etc.). These updates were made based on (1) lessons learnt and pitfalls observed in the measurement procedure during year 1 (updates to mathematical formulae) and (2) needs from the community of practice members to fine grain these criteria to match their expectations. The rest of the document is organised as follows:

Chapter 1 presents an overview of the research outputs chosen for assessment in year 2 along with their description. We have expanded the categorisation schema of these tools after D3.2 to include two new ones – vocabulary mapping and technical metadata extractors. The reason was because the tools chosen in these corresponding categories did not fit in well with the generic assessment criteria i.e. metadata mapping and information extraction, as these tools (Amalgame and MXF tool chain) require specific measurement functions in order to perform tests which fall outside the generic templates.

Chapter 2 presents the updates made to the assessment criteria. The templates defined in D3.1 and D3.2 are reused wherever they can directly applied without modification (e.g. Storage templates). Most of the other categories needed updates in terms of the measurement criteria and the underlying functions.

Chapter 3 presents a detailed assessment results of all the tools chosen in year 2. One thing to note here is that we have also tested two commercial hardware setups as part of this task for storage systems (LTO6 and Optical Drives). The full results of these tests are report in a separate annex and are confidential until we receive approval from the commercial vendors to be publicly disseminated.

Chapter 4 describes the additions made to the Presto4U year 1 dataset based on requirements of the tools chosen for assessment. A combination of open and free to download datasets, and in house produced files (specific to the tool being assessed) proved to be the ideal testing dataset within the project.

Finally we conclude in Chapter 5.

1 Research Outputs Identification

In the context of Presto4U a **Research Output (RO)** is a software/hardware/methodology which is a direct result of research in AV and other digital preservation projects and which has the potential for commercial up-take in the future. We are specifically looking at EC FP6 and FP7 projects (results of PRESTOPRIME and previous Presto family projects are being monitored). As part of the technology and market watch sub task of WP3 task T3.2, we will also look at commercial ROs during the course of the project.

As part of task 3.2, the first stage is in identifying the ROs which can potentially address the CoP needs and present an opportunity for take-up. For the second stage, in order to objectively quantify the suitability of an RO, we need to assess the tool using a formally defined methodology and a measurement method. A measurement method is a logical sequence of operations used to quantify properties with respect to a specified scale. The result is a quality measure element. Therefore, in order to measure software quality we need for each specified characteristic to define:

- measure elements, e.g. identify which set of system properties cover a quality characteristic
- measurement method or test which measures each system property. The combination of those measures will derive the quality measure of that characteristic.

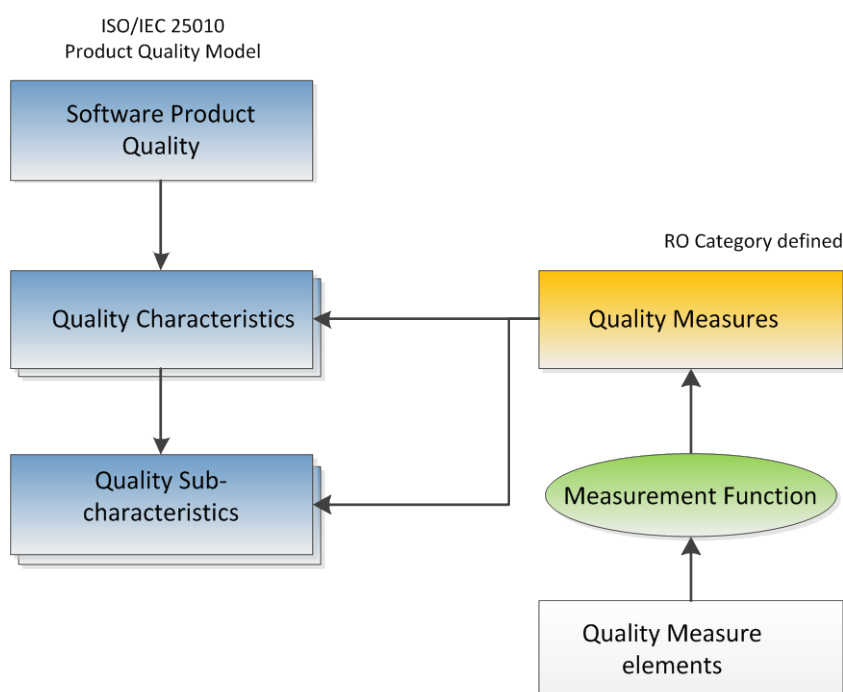


Figure 1 Relations between quality model and measures

Ideally, there should be a clear mapping between user requirements regarding the assessment and the quality characteristics provided by the standard. We call this quality requirements. Such quality requirements can be used to define measure elements as well as the measurement methods. The measurement methods will be applied on the RO during assessment. In order to carry out the assessment exercise, we have firstly specified a general assessment methodology and further specialised the quality measures for each

category of tool which is being evaluated. This allows us to define generic quality measure based on tool categories and also enables the comparison of tools within the same category.

1.1 Research Outputs Chosen

This section will provide a description of the tools chosen as part of Year 2. Two new categories of tools, namely, vocabulary mapping and technical metadata extractors were added in year 2.

1.1.1 Metadata mapping

For metadata mapping tools, the assessment of the tools performed in year 1 has been updated. Descriptions of the tools being re-tests i.e. PrestoPRIME Metadata Mapping Tool, MINT Mapping Tool can be found in D3.2 [4] and will not be repeated again in this deliverable. We start with the description of a vocabulary mapping tool – Amalgame.

1.1.2 Vocabulary mapping

Vocabulary mapping is a related topic to metadata mapping, but deals with the conversion of controlled vocabularies, thesauri and ontologies. It thus requires another type of tools which are assessed according to a specific set of criteria.

1.1.2.1 Amalgame Vocabulary Mapping

Amalgame (AMsterdam ALignment GenerAtion MEtatool) is a tool for finding, evaluating and managing vocabulary alignments. The tool has been developed in the context of the Ontology Alignment Evaluation Initiative (OAEI), in which different alignment methods can be combined using a workflow setup.

The Amalgame Alignment server features:

- A workflow composition functionality, where various alignment generators can be positioned. Their resulting mapping sets can be used as input for filtering methods, other alignment generators or combined into overlap sets.
- A statistics function, where statistics for alignment sets will be shown
- An evaluation tool, where subsets of alignments can be evaluated manually

Amalgame has been used in a variety of use cases including aligning GTAA vocabulary of the Dutch Sound and Vision Institute with Dutch lexical thesaurus Cornetto. WordNet 2.0 and 3.0, GEMET-Agrovoc for multi-lingual metadata mapping, various pilot projects for the Amsterdam Museum including publishing museum metadata as Lined Open Data, PICO-AAT for mapping to the Art and Architecture Thesaurus etc.

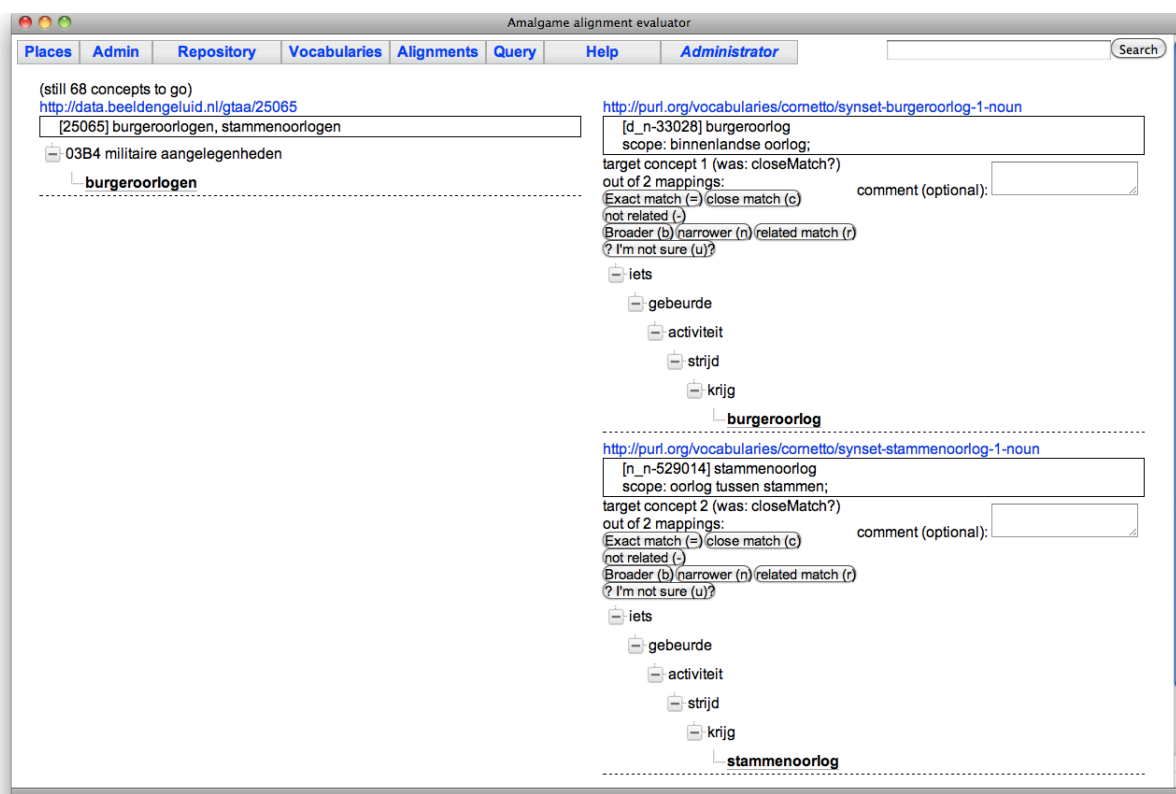


Figure 2: Alamgame Vocabulary Mapping Tool

1.1.3 Quality assessment

The assessment of VidiCert has been updated. The description of this tool was already reported in D3.2. A new set of tools - the BAVC QC Tools have been added to the assessment set in year 2.

1.1.3.1 BAVC QC Tools

QC Tools is an application developed within a project lead by the Bay Area Video Coalition (BAVC)¹ with the purpose of creating new software tools that can report on and graph data documenting video signal loss, flag errors in digitization, identify which errors and artifacts are in original format and which resulted from the digital transfer based on all the data collected in the past. Partners of the project are Dance Heritage Coalition² and independent consultant David Rice. The project is funded by the National Endowment for the Humanities under Grant number PR-50188-13. It was initiated in January 2013 and will last until end of January, 2015.

QC Tools is available as open source and licensed under a GPLv3 license. For the following evaluation, version 0.6.2 of QC Tools for Windows as released on November 3rd 2014 was used. The user interface for QC Tools is shown in Figure 3. Video files can be added to the files list and get immediately analysed by QC Tools. The QC task consists then by inspecting the low-level signal filters visualized as line charts over a timeline by a QC expert.

¹ <http://www.bavc.org/qctools>

² <http://www.danceheritage.org/>

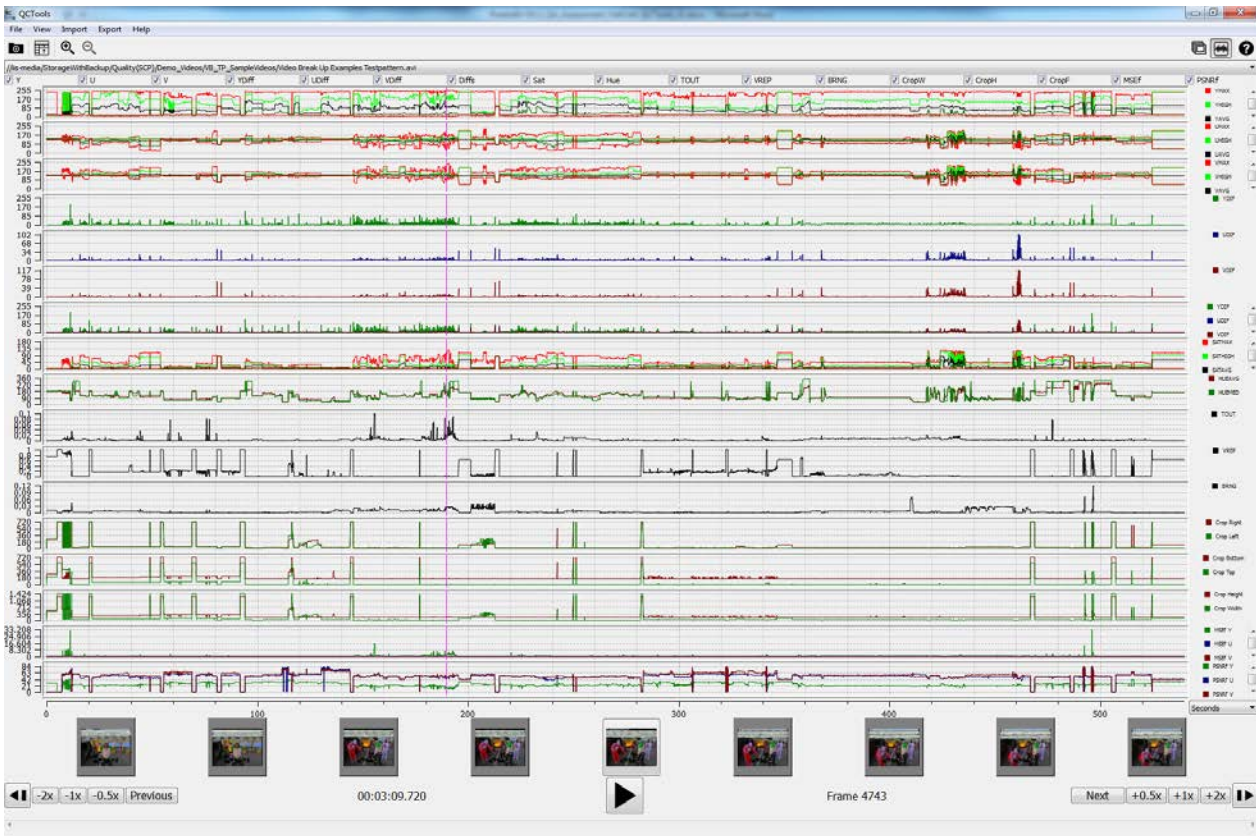


Figure 3: BAVC QC Tools User Interface.

Playback functionality is available in QC Tools with various playback filters shown in Figure 4. Two visualisation components are shown next to each other and can either display the original video image or a selected filter applied on that image.

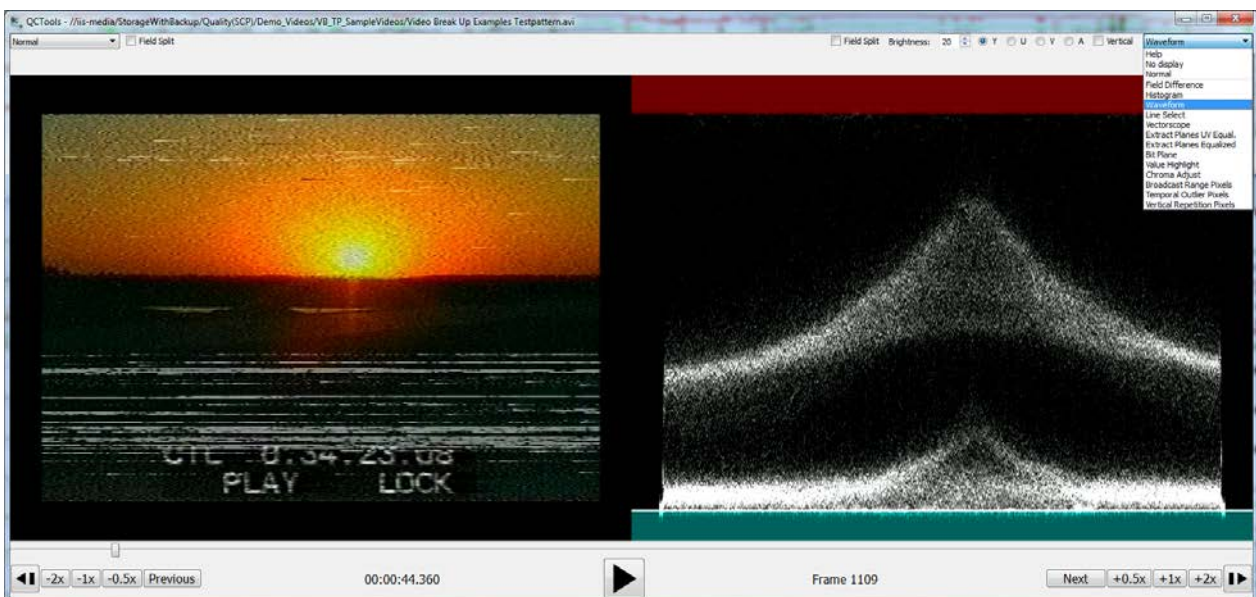


Figure 4: BAVC QC Tools Playback Functionality.

1.1.4 Technical Metadata Extractors

1.1.4.1 MXFDump

Is a command line tool based on open source C++ software MXFLib. MXFDump works with generic MXF files and returns a plain text with a very deep and detailed report of the file structure.

1.1.4.2 MXFAnalyzer

A commercial tool used for the analysis of generic MXF files. The analysis includes compliance check against relevant SMPTE standards and detailed reporting. It can be used with a dedicated GUI, as command line or through web services.

1.1.4.3 MXFTechMDEExtractor

Developed by RAI, it is a software component tool made available as open-source under GPL v3 licence. The tool is written in Java and is made available both as a JAR library and as source in an Eclipse project package. The tool analyzes only the header of generic MXF files and reports the most relevant technical metadata e.g. Operational Pattern, video resolution, aspect ratio etc. Can be used through command line or integrated in a wider Java project.

1.1.4.4 MediaInfo

MediaInfo is an open source software for generic multimedia file analysis written in C++ language. The tool extracts the most relevant technical and tag data for video and audio, giving detailed and configurable reports. A wide range of media formats and coding is supported. It can be used with a dedicated GUI (Windows), as command line or integrated in a wider software project as a library (LDD).

1.1.4.5 FFProbe

FFprobe is a command line tool based on open source software FFmpeg. FFprobe gathers information from multimedia streams and prints it in human- and machine-readable fashion. A wide range of media formats and coding is supported. Despite it accepts MXF files, it is not specialized on that and does not output particular information like the MXF Operational Pattern.

1.1.5 Preservation and Platform Systems

Archivematica has already been described I D3.2. In this section, we describe the new tools for preservation and platforms chosen in year 2.

1.1.5.1 DSpace

DSpace is an out of the box open source repository application for delivering digital content to end-users, typically used for creating open access repositories for scholarly

and/or published digital content. Due to its wide adoption it can be considered one of the most widely used open source repository software for non-profit and commercial organisations.

DSpace captures, stores, indexes, preserves and redistributes an organization's research material in digital formats. Research institutions worldwide use DSpace for a variety of digital archiving needs from institutional repositories (IRs) to learning object repositories or electronic records management, and more. DSpace can be customized and extended.

An active community of developers, researchers and users worldwide contribute to DSpace community. While DSpace shares some feature overlap with content management systems and document management systems, the DSpace³ repository software serves a specific need as a digital archives system, focused on the long-term storage, access and preservation of digital content.

1.1.5.2 RODA

Based upon Fedora, RODA is a complete digital repository that provides functionality for all the main units of the OAIS reference model and it is maintained by KEEP SOLUTIONS⁴. This platform is based on open source technologies and takes advantage of existing standards such as METS, EAD and PREMIS. It is possible to add more functionality to the system by means of a plug-in and task scheduling mechanism. The repository natively supports normalization on ingest for different file formats and migration-based preservation actions.

A task scheduler takes care of preservation actions and management, defining the set of rules that trigger specific actions, and when these should take place. RODA Core Services are responsible for carrying out more complex tasks such as handling the ingest workflow, querying the repository in advanced ways and carrying out administrative functions on the repository. The platform can be integrated with systems already existing in the client institution⁵.

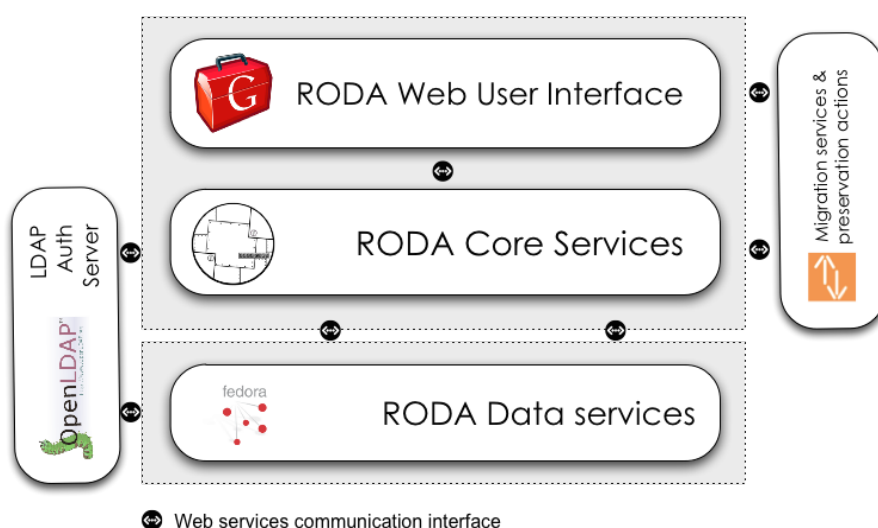


Figure 5: RODA Architecture

³ <http://www.dspace.org>

⁴ <http://www.keep.pt>

⁵ <http://www.roda-community.org>

2 Assessment Criteria Updates

This chapter will present updates to the assessment methodology and assessment templates based on the lessons learnt from Year 1 tool testing and revision and redefinition of the measurement plan that takes more in consideration the peculiarities and exigencies of different CoPs.

2.1.1 Metadata mapping

In this section a measurement plan for the metadata mapping RO category is defined, updating the measurement plan from D3.2. In particular, some functional criteria have been updated, and the calculation of scores has been refined for a number of criteria.

Firstly, we start by defining the functions required to be tested followed by a measurement plan on specific functions which need to be specialised (specialisation from the generalised criteria mentioned in the section above) for this particular category. The levels of need are classified as follows:

- Mandatory - Must have
- Recommended - Could deal also without, but it would be better to have
- Desirable - May be appreciated in some cases, but in most cases it doesn't make the difference

2.1.1.1 Definition of functions

Functions	Levels of Need	Description
<i>Metadata input formats</i>		Support for a metadata format as source format of the mapping process
Dublin Core	Mandatory	
ESE	Desirable	
EDM	Desirable	
EBU Core	Mandatory (for broadcasting related CoPs) / Recommended	
MPEG-7	Recommended (for CoPs using automatic content analysis)	
LIDO	Mandatory (for museum/gallery related) / Recommended	
EAD	Mandatory (for non-a/v archive related) / Recommended	
<i>Metadata output formats</i>		Support for a metadata format as target format of the mapping process
Dublin Core	Mandatory	
ESE	Recommended	
EDM	Recommended	

Functions	Levels of Need	Description
EBU Core	Mandatory (for broadcasting related CoPs) / Recommended	
MPEG-7	Recommended (for CoPs using automatic content analysis)	
LIDO	Mandatory (for museum/gallery related) / Recommended	
EAD	Mandatory (for non-a/v archive related) / Recommended	
option to add custom formats	Recommended	Support for adding new metadata formats
XML representation support	Mandatory	Support for metadata documents in XML format
RDF representation support	Recommended	Support for metadata documents in RDF format
<i>Metadata model constructs</i>		
single -> multiple elements	Recommended	Support mapping a single element into a set of elements (e.g., string into structured)
multiple -> single elements	Mandatory	Support mapping a set of elements into a single elements (e.g., structured into string)
structure using context elements	Mandatory	Define mapping of content structure constructs using contextual elements
conditional mapping based on element/attribute values	Mandatory	Define mapping rules that are conditioned on values of elements or attributes
map collections	Recommended	Support for mapping of collections of metadata records rather than single metadata records only
merge string values	Mandatory	Support merging values of separate string values into one string value
split string values	Recommended	Support splitting a string value into a set of separate string values
number of levels in data structure	Mandatory: 2 / Recommended: 2+	Support for number of hierarchy levels in the document structure
start from example(s)	Recommended	Support initiating mapping from example documents
start from schema	Recommended	Support initiating mapping from a schema instance
configuration user interface	Mandatory	Provision of a configuration user interface (instead of/in addition to configuration files)
<i>user interface</i>		
drag & drop mappings	Recommended	Support of drag&drop for configuring the mappings
preview	Mandatory	Provide preview of configured mappings
map constructs not found in available examples	Recommended	Support the definition of mappings for constructs not found in one of the examples

Table 1: List of functions for metadata mapping

2.1.1.2 Measurement plan

Here below, for metadata mapping RO, the measurement plan has been customized in some characteristics and sub characteristics by their measures. To this aim, peculiarities and specific features have been taken in consideration. In particular some metadata/vocabulary mapping tools are automatic. However, they have a user interface for configuration of the mapping, thus the UI criteria can be applied to it. Vocabulary mapping tools follow similar workflows, and only functional criteria differ significantly.

For general information on the measurements see Section 2 of D3.2.

2.1.1.2.1 *Functional suitability*

Degree to which a product or system provides functions that meet stated and implied needs when used under specified conditions.

Measurement function: $FS = (X+Y+Z) / 3$ (where X,Y,Z are the scores computed as in the following)

FS = Functional Suitability

X = Functional Completeness

Y = Functional Correctness

Z = Functional Appropriateness

In particular the ability to reach a defined goal using the tool can be considered

Interpretation of test results: FS value closer to 1 is better

Functional Completeness:

1. Measure: functional metadata formats coverage
2. Description of measure: Evaluation of the metadata formats supported and the constructs of data model supported. Measured by comparing against widely adopted metadata formats/models.
3. Measurement function: $X = 1/n \sum_{i=1..n} w_i f_i$, with n being the number the functions, w_i the weight of the function (mandatory = 1, recommended = 0.75, desired = 0.5) and f_i the degree to which the function is provided [0..1] (with 1 = completely provided).

Functional Correctness:

1. Measure: functional correctness of mapping
2. Description of the measure: Correctness and completeness of the mapping between a pair of formats.
Note: The evaluation of the mapping can be performed and validated by an expert.
3. Measurement function: $Y = 1/m \sum_{i=1..m} c_i$, with m being the number of mapping checked, and c_i the assessment of the correctness of the mapping (with 1 = fully correct).

Functional Appropriateness:

1. Measure: functional appropriateness of the mapping tools

2. Description: The evaluation considers mappings required in common preservation workflow steps (ingest, B2B exchange, export to Europeana) as assessed by an expert.
3. Measurement function: $Z = 1/p \sum_{i=1..p} a_i$, with p being the number of workflows checked, and a_i the assessment of the appropriateness for the workflow (with 1 = fully appropriate).

2.1.1.2.2 Performance efficiency

Performance relative to the amount of resources used under stated conditions.

Measurement function: $PE = (X+Y+Z) / 3$ where

PE = Performance Efficiency

X = Time behaviour

Y = Resource utilization

Z = Capacity

Interpretation of test results: PE smaller is better

Time behaviour:

Measure: The mean processing time X in milliseconds for performing a defined set of mapping problems.

Interpretation of test results: X varies from 0 to infinite. Smaller is better.

Resource utilization:

Degree to which the amounts and types of resources used by a product or system when performing its functions meet requirements

4. Measure: (Mean) CPU/RAM utilization
5. Description of the measure: How much CPU time/RAM is used to perform a given task
6. Measurement function:
 $Y = 0.5 * (\text{fraction of CPU} + \text{fraction of RAM})$ actually used to perform a task on a reference system
7. Interpretation of test results: Y smaller is better.

Capacity:

Degree to which the maximum limits of a product or system parameter meet requirements

Measure: Maximum throughput using a specific reference configuration

1. Measurement function:
 $Z = A/B$ where
A = operation time
B = the total no. of processed documents
2. Interpretation of test results: Z smaller is better.

2.1.1.2.3 **Compatibility**

Degree to which a product, system or component can exchange information with other products, systems or components, and/or perform its required functions, while sharing the same hardware or software environment.

Measurement function: $Co = Y$ where

Co = Compatibility

Y = Interoperability

Interpretation of test results: Co value larger is better

Interoperability:

Degree to which two or more systems, products or components can exchange information and use the information that has been exchanged

1. Measure: Supported service interfaces for information exchange (metadata files, REST, SOAP)
2. Description of the measure: the service interfaces are smoothly exchanged with other software or systems

3. Measurement function:

$Y = A / B$ where

A = number of interfaces for information exchange

B = total number of interfaces to be supported

4. Interpretation of test results: Y varies from 0 to infinite. Usually, larger is better.

2.1.1.2.4 **Usability**

Degree to which a product or system can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.

Measurement function: $Us = N$, where

Us = Usability

N = Accessibility

Interpretation of test results: Us value closer to 1 is better

Note: Due to resource limitations, no actual user studies can be performed.

Accessibility:

1. Measure: The control elements of the user interface that provide accessibility by adaptable text sizes, alternative image texts and supporting multiple input devices.

2. Measurement function:

$$N = A / B \text{ where}$$

A = number of control elements supporting accessibility options

B = total number of control elements checked

3. Interpretation of test results: N varies from 0 to 1. Larger is better.

2.1.1.2.5 Reliability

Degree to which a system, product or component performs specified functions under specified conditions for a specified period of time.

Measurement function: $Re = L$, where

Re = Reliability

L = Recoverability

Interpretation of test results: Re value closer to 1 is better

Recoverability:

The possibility of using the tool in the following situations:

- incorrect input file format
- load incomplete mapping definition
- user creates contradicting mapping
- wrong drag & drop operation
- example file in wrong format

$L = 1/n \sum_{i=1..n} r_i$, where n is the number of situations tested, and r_i is the assessment of the recoverability in that situation [0..1] (1 = fully recoverable without side effects).

2.1.1.2.6 Maintainability

Degree of effectiveness and efficiency with which a product or system can be modified by the intended maintainers.

Measurement function: $Ma = (H+J) / 2$ where

Ma = Maintainability

H = Modularity

J = Reusability

Interpretation of test results: Ma value closer to 1 is better

Modularity:

1. Measure: Easiness of adding new formats, data types and mapping rules.

2. Measurement function:

H = assessment of effort for adding a new mapping

3. Interpretation of test results: H varies from 0 to 1. Closer to 1 is better.

Reusability:

Degree to which an asset can be used in more than one system, or in building other assets

1. Measure: Integration in other system based on documented interfaces and exchange formats.
2. Description of the measure: Supported service interfaces.
3. Measurement function:

$$J = A / B$$

A = number of service interfaces/exchange formats supported

B = total number of service interfaces/exchange formats considered

Interpretation of test results: J varies from 0 to 1. Usually, closer to 1 is better.

2.1.2 Vocabulary mapping

In this section a measurement plan for vocabulary mapping RO category is defined. Firstly, we start by defining the functions required to be tested followed by a measurement plan on specific functions which need to be specialised (specialisation from the generalised criteria mentioned in the section above) for this particular category. The levels of need are classified as follows:

- Mandatory - Must have
- Recommended - Could deal also without, but it would be better to have
- Desirable - May be appreciated in some cases, but in most cases it doesn't make the difference

2.1.2.1 Definition of functions

Functions	Levels of Need	Description
<i>Vocabulary formats</i>		Support for a vocabulary format
SKOS	Mandatory	
Zthes (Z39.50)	Recommended	
ISO 25964	Recommended	
MPEG-7 CS	Desirable	
Wordnet RDF	Desirable	
MARC-21	Desirable	
option to add custom formats	Recommended	
<i>Mapping options</i>		
term -> term	Mandatory	
broader	Recommended	

Functions	Levels of Need	Description
narrower	Recommended	
related	Mandatory	
automatic mapping	Recommended	
manual/semi-automatic mapping	Mandatory	
<i>user interface</i>		
drag & drop mappings	Recommended	
preview	Mandatory	

Table 2: List of functions for vocabulary mapping

2.1.2.2 Measurement plan

Here below, for vocabulary mapping RO, the measurement plan has been customized in some characteristics and sub characteristics by their measures. To this aim, peculiarities and specific features have been taken in consideration. In particular some metadata/vocabulary mapping tools are automatic. However, they have a user interface for configuration of the mapping, thus the UI criteria shall be applicable to it as well. Vocabulary mapping tools follow similar workflows, and only functional criteria differ significantly.

For general information on the measurements see Section 2 of D3.2.

2.1.2.2.1 *Functional suitability*

Degree to which a product or system provides functions that meet stated and implied needs when used under specified conditions.

Measurement function: $FS = (X+Y+Z) / 3$ (where X, Y, Z are the scores computed as in the following

FS = Functional Suitability

X = Functional Completeness

Y = Functional Correctness

Z = Functional Appropriateness

In particular the ability to reach a defined goal using the tool can be considered

Interpretation of test results: FS value closer to 1 is better

Functional Completeness:

4. Measure: vocabulary formats coverage
5. Description of measure: Evaluation of the vocabulary formats supported and the mapping operations supported.
6. Measurement function: $X = 1/n \sum_{i=1..n} w_i f_i$, with n being the number the functions, w_i the weight of the function (mandatory = 1, recommended = 0.75, desired = 0.5) and f_i the degree to which the function is provided [0..1] (with 1 = completely provided).

Functional Correctness:

4. Measure: correctness of mapping
5. Description of the measure: Correctness and completeness of the mapping between a pair of vocabularies.
Note: The evaluation of the mapping can be performed and validated by an expert.
6. Measurement function: $Y = 1/m \sum_{i=1..m} c_i$, with m being the number of mapping checked, and c_i the assessment of the correctness of the mapping (with 1 = fully correct).

Functional Appropriateness:

8. Measure: functional appropriateness of the tools
9. Note: The evaluation considers vocabulary required in common preservation workflow steps (ingest, B2B exchange, export to Europeana) as assessed by an expert.
10. Measurement function: $Z = 1/p \sum_{i=1..p} a_i$, with p being the number of workflows checked, and a_i the assessment of the appropriateness for the workflow (with 1 = fully appropriate).

2.1.2.2.2 Performance efficiency

Performance relative to the amount of resources used under stated conditions.

Measurement function: $PE = (X+Y+Z) / 3$ where

PE = Performance Efficiency
 X = Time behaviour
 Y = Resource utilization
 Z = Capacity

Interpretation of test results: PE smaller is better

Time behaviour:

Measure: The mean processing time X in milliseconds for performing a defined set of mapping operations.

Interpretation of test results: X varies from 0 to infinite. Smaller is better.

Resource utilization:

Degree to which the amounts and types of resources used by a product or system when performing its functions meet requirements

7. Measure: (Mean) CPU/RAM utilization
8. Description of the measure: How much CPU time/RAM is used to perform a given task
9. Measurement function:

$$Y = 0.5 * (\text{fraction of CPU} + \text{fraction of RAM}) \text{ actually used to perform a task on a reference system}$$
10. Interpretation of test results: Y smaller is better.

Capacity:

Degree to which the maximum limits of a product or system parameter meet requirements

Measure: Maximum throughput using a specific reference configuration

3. Measurement function:

$$Z = A/B \text{ where}$$
 - A = operation time
 - B = the total no. of processed terms
4. Interpretation of test results: Z smaller is better.

2.1.2.2.3 Compatibility

Degree to which a product, system or component can exchange information with other products, systems or components, and/or perform its required functions, while sharing the same hardware or software environment.

Measurement function: $Co = Y$ where

$Co = \text{Compatibility}$
 $Y = \text{Interoperability}$

Interpretation of test results: Co value larger is better

Interoperability:

Degree to which two or more systems, products or components can exchange information and use the information that has been exchanged

5. Measure: Supported service interfaces for information exchange (metadata files, REST, SOAP)
6. Description of the measure: the service interfaces are smoothly exchanged with other software or systems
7. Measurement function:

$$Y = A / B \text{ where}$$
 - A = number of interfaces for information exchange

B = total number of interfaces to be supported

8. Interpretation of test results: Y varies from 0 to infinite. Usually, larger is better.

2.1.2.2.4 Usability

Degree to which a product or system can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.

Measurement function: $U_s = N$, where

U_s = Usability

N = Accessibility

Interpretation of test results: U_s value closer to 1 is better

Note: Due to resource limitations, no actual user studies can be performed.

Accessibility:

4. Measure: The control elements of the user interface that provide accessibility by adaptable text sizes, alternative image texts and supporting multiple input devices.

5. Measurement function:

$N = A / B$ where

A = number of control elements supporting accessibility options

B = total number of control elements checked

6. Interpretation of test results: N varies from 0 to 1. Larger is better.

2.1.2.2.5 Reliability

Degree to which a system, product or component performs specified functions under specified conditions for a specified period of time.

Measurement function: $Re = L$, where

Re = Reliability

L = Recoverability

Interpretation of test results: Re value closer to 1 is better

Recoverability:

The possibility of using the tool in the following situations:

- incorrect input file format
- load incomplete mapping definition
- user creates contradicting mapping
- wrong drag & drop operation
- example file in wrong format

2.1.2.2.6 **Maintainability**

Degree of effectiveness and efficiency with which a product or system can be modified by the intended maintainers.

Measurement function: $Ma = (H+J) / 2$ where

Ma = Maintainability

H = Modularity

J = Reusability

Interpretation of test results: Ma value closer to 1 is better

Modularity:

4. Measure: Easiness of adding new vocabularies and mapping rules.

5. Measurement function:

H = assessment of effort for adding a new mapping

6. Interpretation of test results: H varies from 0 to 1. Closer to 1 is better.

Reusability:

Degree to which an asset can be used in more than one system, or in building other assets

4. Measure: Integration in other system based on documented interfaces and exchange formats.

5. Description of the measure: Supported service interfaces.

6. Measurement function:

$J = A / B$

A = number of service interfaces/exchange formats supported

B = total number of service interfaces/exchange formats considered

Interpretation of test results: J varies from 0 to 1. Usually, closer to 1 is better.

2.1.3 **Quality Assessment**

In this section a measurement plan for audiovisual quality assessment RO category is defined, updating the measurement plan from D3.2. In particular, some functional criteria have been updated, and the calculation of scores has been refined for a number of criteria. Firstly, we start by defining the functions required to be tested followed by a measurement plan on specific functions which need to be specialised (specialisation from the generalised criteria mentioned in the section above) for this particular category. The levels of need are classified as follows:

- Mandatory - Must have
- Recommended - Could deal also without, but it would be better to have
- Desirable - May be appreciated in some cases, but in most cases it doesn't make the difference

2.1.3.1 Definition of functions

Functionality	Level of need	Description
Automatic Defect Analysis Functions (categorisation based on EBU QC checks ⁶)		
Analogue Synchronisation Errors Aliases: lost lock, time-base corrector (TBC) hit, video breakup, lost video sync, horizontal distortion	Mandatory	System shall check for analogue synchronisation problems that have caused severe visual line/field/frame distortions. Analogue synchronisation problems can lead to visible artefacts. These can be created during the analogue tape read process (e.g. during tape digitisation), as part of the analogue video transmission process or as a side-effect of improper analogue video editing. It manifests itself in visual line/field/frame/multi-frame distortions of varying degree with typically a horizontal/line oriented appearance and a temporal extent of one or more fields/frames. Problems with vertical synchronization usually result in rolling (up or down) frames. As a severity measure the relation between the distorted area and the area of the entire frame is used.
Coloured Frames Aliases: Black Frames, Monochrome Frames, Uniform Color Frames	Mandatory	System shall detect frames which have no active video and point out full-sized single coloured frames. Coloured frames may be produced by video tape players (during migration) or by software errors in the production cycle.
Digital Tape Dropouts Aliases: digital video tape dropout, digital hits, digital tape hits	Mandatory	System shall detect visible artefacts caused by digital tape errors. The result may include tape/error type and severity together with spatial locations. This is about visible artefacts which occur within the digital tape read process and manifests itself when head problems or tape overuse cause the error correction of the VTR to create short term failures of parts of frames. The visual effect is the appearance of impairments, such as alternating lines in a block, duplicated block areas, arrays of similar pixels within a block area, and random portions of blocks with changed luminance or chrominance within one or multiple consecutive frames. The appearance of those blocks as well as the spatiotemporal pattern of those blocks strongly depends on the kind of tape, such as DigiBETA, IMX, DV.... Most relevant in the context of archive migration applications are early digital tape formats, e.g. DigiBETA.
Video Noise Aliases: image noise, noise	Desirable	System shall detect video segments whose essence shows a noise level that is above a user-defined threshold. The visual noise level might be estimated by a signal to noise ratio (SNR). Noise constitutes an unwanted signal that inevitably adds to the useful part, it may originate from different sources, e.g. electronic sensor noise, quantisation noise, film grain noise.... For archive applications the knowledge on the noise level is relevant to estimate restoration costs for content re-used (e.g. in a program, DVD, BD...)
Blurriness Aliases: out of focus, blur detection, sharpness	Recommended	System shall detect video segments whose image content would be perceived as blurry by the viewer. For archive applications the knowledge on content blurriness is relevant to decide if it can be re-used for a certain purpose (e.g. is SD content sharp enough to be re-used for an HD program, BD,...)

⁶ EBU Strategic Programme on QC (EBU QC) <http://tech.ebu.ch/groups/qc>, First draft release of QC test definitions available at <http://tech.ebu.ch/docs/tech/tech3363.zip>

Functionality	Level of need	Description
Video Test Pattern Aliases: test card, colour bars	Mandatory	System shall detect video segments containing specific test pattern content. A test pattern is a sequence of (often still) images with showing particular characteristics. For video experts, test patterns allow to quickly detect problems in a generic video chain and facilitate calibration, alignment, and matching of video devices. In typical broadcaster workflows, test patterns often have to be cut off or checked for a specific position and duration(e.g. at the beginning and end of a programme). For archive applications test pattern segments shall be detected after the migration of content /programs, especially on multi-program tapes. Usually no test pattern segments shall be present in a file containing a single program.
Video Field Order Aliases: field order, field dominance	Mandatory	System shall detect video segments containing a field order differing from an expected one.
Scanning Type Aliases: sampling, sampling structure, scanning	Mandatory	System shall detect video segments containing a scanning type different from an expected one, e.g. interlaced, progressive or pull-down
Audio Silence Aliases: mute test, minimum level	Mandatory	System shall check if the audio level is lower than a user defined threshold value. In archive migration applications the actual audio channel usage can be assessed by audio silence detection. The actual audio channel usage in the video needs to correspond with the audio channel usage described in a content/asset management system.
Audio Encoding Format Change	Recommended	System shall check if the audio encoding is changing within a channel of a program, e.g. from PCM to Dolby-E. The actual audio encoding used in the video for the individual channels needs to correspond with the audio encoding format described for these channels within a content/asset management system.
General Analysis Properties		
Analysis profiles	Mandatory	Capability to adapt quality analysis functions (detectors) and its parameters to the desired QC task/use case
No reference video required	Mandatory	For content within archives stored or to be migrated very often only one copy do exist. The capability to assess the audiovisual quality without any other copy required is therefore crucial
Detection of multi-generation defects	Recommended	Defects within content of AV archives may have been copied/migrated from one earlier format to the next. These defects (e.g. analogue synchronisation errors or analogue tape dropouts) are visible within the current copy on a certain media format (e.g. DigiBETA) but are not originated from the current format or encoding. An AV quality assessment system for AV archive assessment or migration shall be able to work independently from the current media format and encoding
Multi-Resolution support	Mandatory	Capability to process content with different nominal resolution, e.g. SD, HD, 2k. Practically any archive holds content with different nominal resolution
GPU support	Recommended	Capability to use the graphics processing unit (GPU) for compute intensive calculations within defect detectors.
Video standard support	Recommended	Capability to read/analyse video provided in a non-proprietary, internationally standardised format e.g. by MPEG or SMPTE

Functionality	Level of need	Description
Metadata standard support	Recommended	Capability to write/provide quality assessment metadata in a non-proprietary, internationally standardised format e.g. by MPEG or SMPTE
Interactive Validation/Verification Functions		
Check file efficiently for correct content	Mandatory	Check efficiently that file contains correct content (potentially described in an content/asset management system) and that correct tape segment has been digitised, e.g. containing no test patterns pre-recorded on multi-program tapes.
Human validation of automatic analysis functions	Mandatory	The system shall support the human verification of detections from an automatic analysis step. In this way verified reports can be generated.
Interactive defect annotation support	Mandatory	Defects can be manually created and modified (time and duration) by a human operator. Detections missed by the automatic analysis can be annotated this way.
Overall quality rating support	Mandatory	The system shall support to give an overall quality rating for the entire content/program (e.g. OK, Error...).
Defect severity based operation/validation	Recommended	The system should support efficient verification by prioritizing the most relevant annotations
Video output devices	Desirable	The system should support the output of videos on the following devices: - Desktop within a single screen - Desktop on a second full screen - SDI
Individual field output	Recommended	The system supports to output individual fields in the GUI
Video output on interlaced capable devices	Desirable	The system supports to output video on interlaced video capable devices
Human validation during analysis phase	Desirable	The system supports to display analysis results as they are detected

Table 3: List of functions for quality assessment ROs

2.1.3.2 Measurement plan

Here below, for Quality Assessment RO, the measurement plan has been customized in some characteristics and sub characteristics by their measures. To this aim, peculiarities and specific features have been taken in consideration.

In particular, we consider automatic and semi-automatic tools, i.e. some of the tools have a user interface, while others are automatic services. Thus, some criteria (most notably the usability related ones), apply only to tools/components with a UI.

For general information on the measurements see Section 2 of D3.2.

2.1.3.2.1 Functional suitability

Degree to which a product or system provides functions that meet stated and implied needs when used under specified conditions

Measurement function: $FS=(X+Y)/2$ (where X and Y are the scores computed as in the following

FS = Functional Suitability

X = Functional Completeness

Y = Functional Appropriateness

Interpretation of test results: FS value closer to 1 is better

Functional Completeness

11. Measure: functional coverage
12. Description of measure: Evaluation of the actual functions supported from the list of all functions (Automatic Defect Analysis Functions, General Analysis Properties and Interactive Validation/Verification Functions) described in section 2.1.3.1 .
Measured by comparing actual functions supported against the full list of functions.
13. Measurement function: $X = 1/n \sum_{i=1..n} w_i f_i$, with n being the number the functions, w_i the weight of the function (mandatory = 1, recommended = 0.75, desired = 0.5) and f_i the degree to which the function is provided [0..1] (with 1 = completely provided).

Functional Appropriateness

11. Measure: functional appropriateness of the audiovisual quality assessment tool
12. Description: The evaluation considers QA functions required in common archive workflow steps (digitisation/migration, ingest, search/selection) as assessed by an expert.
13. Measurement function: $Z = 1/p \sum_{i=1..p} a_i$, with p being the number of workflows checked, and a_i the assessment of the appropriateness for the workflow (with 1 = fully appropriate).

2.1.3.2.2 Performance efficiency

Performance relative to the amount of resources used under stated conditions.

Measurement function: $PE = (X+Y+Z) / 3$ where X,Y,Z are the scores computed as in the following

PE = Performance Efficiency
 X = Time behaviour
 Y = Resource utilization
 Z = Capacity

Interpretation of test results: PE smaller is better

Time behaviour

Degree to which the response and processing times and throughput rates of a product or system, when performing its functions, meet requirements.

Measures: *Automatic Analysis Time, UI Response Time*

Measure: *Automatic Analysis Time*

Description of the measure: How much time is required for the automatic analysis when performing a given task, e.g. analysis of SD content with a certain video duration.

Measure: *UI Response Time*

Description of the measure: Average User Interface Response Time in seconds for a set of functions provided by the user interface assessed by an expert.

Measurement function:

$$X = \text{Automatic Analysis Time} / \text{duration of video} + \text{average UI Response Time}$$

Interpretation of test results: X smaller is better.

Resource utilization:

Degree to which the amounts and types of resources used by a product or system when performing its functions meet requirements

Measure: (Mean) CPU/RAM utilization

Description of the measure: How much CPU / RAM is actually used to perform a given task

Measurement function:

$$Y = \text{fraction of CPU} + \text{fraction of RAM actually used to perform an task on a reference system}$$

Interpretation of test results: Y smaller is better.

Capacity:

Degree to which the maximum limits of a product or system parameter meet requirements

Measure: Capability to scale analysis throughput

Description of the measure : The system shall be able to scale throughput by the following methods:

C1: Activate/Deactivate detectors based on the customers' needs

C2: Configure and optimize parameters on customers' contents

C3: Scale with additional hardware resources like more CPU cores

C4: Scale with the ability to distribute analysis on multiple computers on a file basis

When a capability (C1 to C4) is fulfilled, its value is 1, otherwise 0.

$$Z = 1/(C1+C2+C3+C4), \text{ for the case C1 to C4 are all 0 } Z=2$$

Interpretation of test results: Z smaller is better.

2.1.3.2.3 Compatibility

Degree to which a product, system or component can exchange information with other products, systems or components, and/or perform its required functions, while sharing the same hardware or software environment.

Measurement function: $Co=Y$ is computed as in the following

$Co = \text{Compatibility}$

$Y = \text{Interoperability}$

Interpretation of test results: Co value larger is better

Interoperability:

Degree to which two or more systems, products or components can exchange information and use the information that has been exchanged

Measure: Interface Support (standardised video input formats, metadata output formats and service interfaces)

Description of the measure: The fraction of standardised video input formats that can be read/analysed by the quality assessment system, metadata output formats that can be written/exported from the quality assessment system and service/workflow interfaces that are supported by the system

Measurement function:

$$Co = Y = A / B \text{ where}$$

A = number of input formats, output formats and interfaces supported

B = total number of input formats, output formats and interfaces, see Table 4.

Interpretation of test results: Y value closer to 1 is better.

Note: The calculation of A is based on the documentation of the test candidates (e.g. information on which container and encoding formats are supported) and if sample files are available they will be used for a quick compatibility check.

Video container formats	The following container formats should be supported: <ul style="list-style-type: none"> - MPEG TS and PS - MXF - MP4 - MOV - AVI
Video encoding formats – often used	The following video formats should be supported: <ul style="list-style-type: none"> - MPEG-2 (incl. IMX50 and XDCAM HD) (e.g. P4U RAI in MXF) - MPEG-4 AVC (H.264) (e.g. P4U UIBK) - JPEG2000 (SAMMA format) - DV and DVCPro
Video encoding formats – less used	The following video formats could be supported: <ul style="list-style-type: none"> - Uncompressed 8/10bit in MOV (e.g. P4U Tate) - Uncompressed 8/10bit in MXF (e.g. BBC) - Uncompressed 8/10bit in AVI - ProRes - DCP, MAP - WMV
Audio encoding formats	The following audio encoding formats should be supported: <ul style="list-style-type: none"> - PCM - MPEG-1 Audio, MPEG-1 Layer 3, MPEG-2 Audio - AAC - AC3 - WMA
Metadata output formats	The output metadata format shall conform to an international metadata standard, e.g. MPEG, SMPTE <ul style="list-style-type: none"> - MPEG-7 - XML
Service/workflow interfaces	The system shall support following service / workflow interfaces for integration: <ul style="list-style-type: none"> - Drop folder - Web service (REST or SOAP)

Table 4: Video input formats (containers and encodings), metadata output formats and service interfaces ideally supported by a quality assessment system.

2.1.3.2.4 Usability

Degree to which a product or system can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.

Measurement function: $Us = N$, where

Us = Usability

N = Accessibility

For calculating the N score, see section below.

Interpretation of test results: Us value closer to 1 is better

Note: Due to resource limitations, no actual user studies can be performed.

Accessibility

Measure: The control elements of the user interface that provide accessibility by adaptable text sizes, alternative image texts and supporting multiple input devices.

Measurement function:

$$N = A / B \text{ where}$$

A = number of control elements supporting accessibility options

B = total number of control elements checked

Interpretation of test results: N varies from 0 to 1. Value closer to 1 is better.

2.1.3.2.5 Reliability

Degree to which a system, product or component performs specified functions under specified conditions for a specified period of time.

Measurement function: $Re = L$, where

Re = Reliability

L = Recoverability

For calculating the L score, see section below.

Interpretation of test results: Re value closer to 1 is better

Recoverability

The possibility of using the tool in the following situations:

- incorrect input file format
- wrong user input during interactive verification, undo
- closing tool/application/system without warning of loss
- network interruption during single file analysis
- system/tool/application termination during multi-file (job) analysis
- history of operator decisions required for current decision

$L = 1/n \sum_{i=1..n} r_i$, where n is the number of situations tested (from the list of situations above), and r_i is the assessment of the recoverability of a specific situation [0..1] (1 = fully recoverable without side effects).

2.1.3.2.6 Maintainability

Degree of effectiveness and efficiency with which a product or system can be modified by the intended maintainers.

Measurement function: $Ma = H$ where

Ma = Maintainability

H = Modularity

For calculating the H score, see section below

Interpretation of test results: Ma value closer to 1 is better

Modularity

Measure: Modular functional extension capability

Measure description: The fraction of functionalities which can be extended in a modular way, desired functionalities are:

- New input video format (wrapper, encoding)
- New defect detector/analysis functionality
- New defect descriptor in output metadata format
- New defect visualisation for interactive verification

Measurement function:

$$H = A / B \text{ where}$$

A = number of system functionalities, which can be extended in a modular way

B = total number of desired modular functionalities, see list above

Interpretation of test results: H varies from 0 to 1. Closer to 1 is better.

2.1.3.2.7 Portability

Degree of effectiveness and efficiency with which a system, product or component can be transferred from one hardware, software or other operational or usage environment to another.

Measurement function: $Po = Y$ where Y computed as in the following

$Po = \text{Portability}$

$Y = \text{Installability}$

Interpretation of test results: Po value closer to 1 is better

Installability

Degree of effectiveness and efficiency with which a product or system can be successfully installed and/or uninstalled in a specified environment

Measure: Ease of installation and availability of documentation

Measure description: Presence of installation functionalities and availability of documentation for the tested system, desired items (installation functionalities and documentation) are:

- Automatic installation tools (wizards)

- Uninstall capability
- Documentation: User manual, Interfaces

Measurement function:

$$Y = A / B \text{ where}$$

A = number of system installation/documentation items supported

B = total number of desired installation/documentation items, see list above

Interpretation of test results: Y closer to 1 is better.

2.1.4 Technical Metadata Extraction Template

The “Technical Metadata Extractors” category has been defined to better tailor the tests and evaluations to be done with a specific family of tools. This category is partly under the more generic “Automatic metadata extraction” one and partly under the “Quality control”, in fact the inspection of technical parameter can be considered as a quality check for assuring file compatibility and usability. In order to belong to the “Technical Metadata Extraction” category, the software tool shall be capable to inspect multimedia files and to extract through simple read-out (or with easy calculation), the principal technical metadata written inside the file itself.

2.1.4.1 Multimedia File Layers

Multimedia files are organized in a sort of matryoshka structure where it is possible to identify three layers: the wrapper, the bitstream and the essence.

The essence is the most inner layer and is constituted by the audio and video samples i.e. the actual content to be played-out, this layer does not contain technical metadata.

The wrapper is the most outer layer, it wraps the video and audio tracks supplying also very important metadata describing the technical characteristics of the content and providing mechanisms like index tables for fast access.

The bitstream layer is the encoded essence, either video or audio. This level contains also technical metadata like the *bitrate* or the *sample rate*.

Often the same metadata item is repeated in the wrapper and in the bitstream.

The distinction into layers leads to identify three possible ways of file inspection for discovering technical metadata:

1. Wrapper inspection

The tool is capable to read-out or simply infer the metadata written inside the wrapper. Examples: the extraction from the wrapper of the declared *video frame rate*, the *display aspect ratio* or the *format* of the enclosed video and audio coding .

2. Bitstream inspection

The tool is capable to read or simply infer metadata declared at bitstream level.

Example: the extraction of declared *display aspect ratio* from the MPEG2 video elementary stream enclosed in a MXF file (MXF is the wrapper and MPEG2 the bitstream).

3. Crosscheck

This is the ability of the tool, to crosscheck the coherence between the wrapper and the bitstream e.g. is the *display aspect ratio* declared in the wrapper equal to what is declared in the bitstream. Crosscheck sometimes applies also within the wrapper when information is in some way repeated, for example the file duration can be written in more places in a MXF wrapper.

Some tools are able to work only with one of the three modalities mentioned e.g. only wrapper, but it is of course preferable to cover all of them.

2.1.4.2 Definition of functions

There exist a core set of technical metadata that applies to the majority of media file formats while other metadata items only make sense for specific formats. MXF is a particular case where metadata is very articulated and sometimes follow a different terminology.

Table 1 lists the most important and common technical metadata items in media files while Table 2 is specific for MXF. The column “MXF name” tells – when different - the specific terminology used by MXF standard documents and “Aliases” the most used alternative names for the same metadata item.

Metadata Item	MXF name	Aliases	Examples	Description
Overall bitrate mode		Bitrate mode	constant variable	It can be constant or variable depending on constant or variable video and audio bitrates.
Overall bitrate		Average overall bitrate	63 Mbit/s 50 Mbit/s	Sum of video and audio bitrates. If the bitrate is variable it is the average.
Video resolution	Stored/Sampled/Displayed height and width (see table 2)	Video height and width, Frame size	1920x1080 720x576	The resolution in pixels to be displayed, given with height and width.
Frame rate mode			constant variable	Can be constant or variable. Broadcast formats are usually constant frame rate.
Video frame rate	Video Sample Rate	Frame rate Video time base Frames per second	25 fps	The frequency to which the video frames are to be presented for a correct playout
Video bit depth	Component depth	Picture bit depth Bit depth Bits per sample	8 bits 10 bits	The number of bits used to quantize each color components of the video signal
Display Aspect ratio	Display Aspect ratio	Aspect ratio AR DAR	16:9 4:3	The horizontal to vertical aspect ratio of the whole image as it is to be presented to avoid geometric distortion
Pixel Aspect ratio		PAR	16:15 1:1	The horizontal to vertical aspect ratio of the single pixel as it is to be presented to reproduce the correct display aspect ratio
Video Scanning	FrameLayout	Scan type Scanning mode Picture scanning Interlacing Interlace mode Frame structure	Progressive Interlaced	Whether the frame is to be scanned progressively or with separated fields (odd and even lines)
Field order	FieldDominance		Upper first Lower first	Which of the field has to be displayed first
Video coding	Picture Essence Coding	Video codec Video format	MPEG2	The algorithm used for compression of the video essence
Video bitrate mode			constant variable	It can be constant or variable
Video bitrate		Video data rate	50 Mbit/s	The bit data rate used to represent the video essence. If variable it is the

				average.
GOP structure			Intraframe	
Color Space		Color Model	YUV, RGB	The way of expressing the possible colors.
Chroma subsampling	HorizontalSubsampling and VerticalSubsampling	chroma profile chroma sampling pixel subsampling	4:2:2	How much is the chrominance subsampled with respect of luminance
Video duration		Duration Playtime Run time	01:00:00	The duration in time units of the video content
Timecode		TC		As defined by SMPTE 12M standard
Audio nbr of channels	Channel Count			The number of separated audio streams
Audio sample rate	Audio Sampling Rate	Audio Sampling frequency	48 Khz	The sampling frequency of the audio signal
Audio bit depth	Audio Quantization bits	Sample depth Bits per sample Audio sample size	24 bits	The number of bits used to quantize the audio sample
Audio coding	Sound Essence Coding	Audio codec Audio format	AES	The algorithm used for compression of the audio essence
Audio bitrate mode			constant or variable	It can be constant or variable
Audio bitrate		Audio rate Audio Data rate	1152 Kbit/s	The bit data rate used to represent the video essence. If variable it is the average.
Audio duration	Audio Track Duration		01:00:00	The duration in time units of the audio content

Table 5 - Common technical metadata

Metadata item	Aliases	Examples	Description
Operational pattern	OP	Op1a	Is a specific way of controlling the complexity of MXF files with respect of contained sources and payout composition
ActiveFormatDescriptor	AFD	AFD 9 (full 4:3 in a 4:3; Pillarbox 4:3 in a 16:9)	Information used for framing the content when the aspect ratio of the display device is different than the aspect ratio of the content e.g. 4:3 in a 16:9
Header partition status		Open and Incomplete Open and Complete Closed and Incomplete Closed and Complete	Open/Closed indicate whether the required header metadata are provisional (possible incorrect values e.g. duration) or final. Complete/Incomplete indicate whether all the best effort header metadata are provisional or final.
Footer partition status		Open and Incomplete Open and Complete Close and Incomplete Closed and Complete	Same as for header partition.
Essence container mapping		MPEG ES Mapping	The essence actually hold inside the container
Stored width and height	Stored frame size, Stored resolution	720x608	Height and width of the stored video frame (what is stored in the file)
Sampled width and height	Sampled frame size, Sampled resolution	720x608	Height and width of the part of the stored rectangle containing only the digital data derived from an image source
Display width and height	Display frame size, Display resolution	720x576	Height and width of the part of the sampled rectangle intended to be displayed

Table 6 - MXF specific technical metadata

2.1.4.3 Measurement Plan

This chapter clearly specifies how to score a tool belonging to this category, Table 3 reports in summary the methodology being used while the following sub-chapters explain the details.

Characteristics	Characteristic Measurement Function	Sub-characteristics	Sub-characteristics Measurement Function
Functional Suitability (FS)	FS = (X + Y) / 2	Functional Completeness (X)	$X = \sum_{i=1}^n Pi/n, Pi \in \{0,1\}$
		Functional Correctness (Y)	$Y = \sum_{i=1}^n Ci/n; Ci \in \{0,1\}$ $Ci = \sum_{j=1}^k Cij/k; Cij \in \{0,1\}$
Performance Efficiency (PE)	PE = (X + Y) / 2	Time Behavior (X)	X = (B - A) / C B-A duration of metadata extraction C maximum allowable duration of extraction
		Resource Utilization (Y)	Y = (A + B) / 2 A CPU percentage used when analyzing a file. B memory perc. used when analyzing a file.
Compatibility (Co)	Co = Y	Interoperability (Y)	Y = (A + B) / 2 $A = \sum_{i=1}^n ai; ai \in \{0,0.5,1\}$ $B = (S + W + L) / 3$
Usability (Us)	Us = (K+L+M)/3	Operability (K)	Y = (A + B) / 2 A,B ∈ {0,0.5,1}
		User error protection (L)	Y = (A + B) / 2 A,B ∈ {0,0.5,1}
		User interface aesthetics (M)	$K = \frac{A + B + C + D + E + F}{6};$ $A, B, C, D, E, F \in \{0,0.5,1\}$
Reliability (Re)	Re = H	Maturity (H)	From 4 to 9 according to the TRL
Maintainability (Ma)	Ma = L	Modifiability (L)	Y = (A + B) / 2 A,B ∈ {0,0.5,1}
Portability (Po)	Po = Y	Installability (Y)	Y = (A + B) / 2 A,B ∈ {0,0.5,1}

Table 7- Measurement plan summary

2.1.4.3.1 Functional Suitability

Degree to which the tool provides functions that meet stated and implied needs when used under specified conditions.

Measurement function: FS=(X + Y) / 2 where

FS = Functional Suitability

X = Functional Completeness

Y = Functional Correctness

Interpretation of test results: FS value closer to 1 is better

Functional completeness: degree to which the set of functions covers all the specified tasks and user objectives.

Will be calculated according to Table 1 and 2 as a count on which metadata item is provided in output or not by the tool. The following formula can be used:

$$X = \sum_{i=1}^n P_i, P_i \in \{0,1\}$$

where “i” is the iterator over the functions indicated in Table1 and Table2, P_i is a boolean number assuming the value 1 when the function is provided (the specific metadata item is extracted) and the value 0 when not provided (metadata item not treated).

Functional correctness: degree to which a product or system provides the correct results with the needed degree of precision.

Will be evaluated for each metadata item whether it is extracted correctly or not and in which percentage on the tested files. The following formula can be used:

$$Y = \sum_{i=1}^n C_i/n; C_i \in [0,1]$$

Where “i” is the iterator over the functions indicated in Table1 and Table2, C_i is a decimal number assuming the values in the interval [0,1], indicating the level of correctness with respect of that function and calculated according to the following formula:

$$C_i = \sum_{j=1}^k C_{ij}/k; C_{ij} \in \{0,1\}$$

where “j” is the iterator over the files used for the evaluation, C_{ij} is a boolean assuming the value 1 when the result of the extraction is correct and 0 when it is wrong.

2.1.4.3.2 Performance Efficiency

Performance relative to the amount of resources used under stated conditions.

Measurement function: $PE = (X+Y) / 2$ where

PE = Performance Efficiency

$X = \text{Time behavior}$

$Y = \text{Resource utilization}$

Interpretation of test results: PE smaller is better

Time behaviour: degree to which the response and processing times and throughput rates of a product or system, when performing its functions, meet requirements.

Time performances will be measured on a reference system with a fixed and specific hardware, operative system, central memory, disks type and configuration. The time behaviour performance will be calculated according to the formula:

$$X = (B - A) / C$$

where B-A is the duration of the metadata extraction operation and C is the maximum allowable duration of that operation. In order to provide a reasonable value to C, we consider the time used to read the entire file on the reference system multiplied by a coefficient k expressing a maximum acceptable overhead.

For example if for checking at bitstream level an MXF/MPEG2 file of one hour video, it takes 6 minutes and for just reading it entirely takes 5 minutes, provided that we use $k=1.5$, the time behaviour score would be $6/(5*1.5)=0.8$

Resource utilization: degree to which the amounts and types of resources used by a product or system when performing its functions meet requirements.

Resource utilization will be shown on the execution timeline through graphs for CPU and memory occupation. The overall resource utilization score will be calculate with the formula

$$Y = (A + B) / 2$$

Where:

A is the average percentage of CPU used by the tool when analyzing a file.

B is the average percentage of memory used by the tool when analyzing a file.

2.1.4.3.3 **Compatibility**

Degree to which a product, system or component can exchange information with other products, systems or components, and/or perform its required functions, while sharing the same hardware or software environment.

Measurement function: $Co = Y$

$Co =$ Compatibility

$Y =$ *Interoperability*

Interpretation of test results: Co value larger is better

Interoperability: degree to which two or more systems, products or components can exchange information and use the information that has been exchanged.

Will be considered two main aspects to be combined according to formula:

$$Y = (A + B) / 2 \quad \text{where:}$$

“A” takes into consideration how much the used terminology for each metadata item in output is common. Despite there is not today a well established standard for the representation and the naming of that technical metadata, each of them has a closed set of well known aliases commonly used. A score of 1 is given when the used term is very common and well known, 0 if unusual and misleading, 0.5 in ambiguous cases.

$$A = \sum_{i=1}^n a_i; a_i \in \{0,0.5,1\}$$

where “i” is the iterator over the extracted technical metadata and “ai” the adequacy of the term being used.

“B” takes into consideration how the tool can be integrated within a wider software environment either with one or more of these modalities: system call (command line), web services (or REST), language library.

$$B = (S + W + L)/3$$

Where S, W and L are booleans assuming the value 1 when system call, web service or library is provided respectively and 0 otherwise.

2.1.4.3.4 Usability

Degree to which a product or system can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.

Measurement function: $U_s = (K + L + M) / 3$ where

U_s = Usability

K = Operability

L = User error protection

M = User interface aesthetics

Interpretation of test results: U_s value closer to 1 is better

Operability: degree to which a product or system has attributes that make it easy to operate and control.

Will be considered the interactive use with either the command line interface or the GUI when available with the following formula:

$$K = \frac{A + B}{2}; A, B \in \{0,0.5,1\}$$

Where:

“A” takes into account whether there an integrated help, 0 if does not exist, 0.5 if exists but not precise and no examples are provided;

“B” takes into account whether export functionality is available (save as xml or other formats)

User error protection: degree to which a system protects users against making errors.

Will be considered the interactive use with either the command line interface or the GUI when available with the following formula:

$$K = \frac{A + B}{2}; A, B \in \{0,0.5,1\}$$

where:

“A” takes into account whether the GUI uses where applicable, controlled vocabularies for the fields. A score of 0 is assigned if it does not exist, 0.5 if exists but not for all the fields, and 1 if always used when possible;

“B” takes into account whether there is a formal check of the input parameter values or in alternative a precise reporting of the problem with inputs, 0 if does not exist, 0.5 if exists but is not precise or not complete, 1 if it exists and is fully satisfactory.

User interface aesthetics:

Degree to which a user interface enables pleasing and satisfying interaction for the user.

The following aspects will be taken into consideration for the evaluation of the interface:

$$K = \frac{A + B + C + D + E + F}{6}; \quad A, B, C, D, E, F \in \{0, 0.5, 1\}$$

- A: Language configurability
- B: Color configurability
- C: Customization of the disposition of fields
- D: Input section completeness
- E: Output display configuration
- F: Presence of an integrated ‘help’

For all of them the score is 1 when present and satisfactory, 0.5 when present but not complete or fully satisfactory, 0 if not present at all.

2.1.4.3.5 Reliability

Degree to which a system, product or component performs specified functions under specified conditions for a specified period of time.

Measurement function: $Re = H$ where

Re = Reliability

H = Maturity

Interpretation of test results: The higher the better

Maturity: degree to which a system meets the needs for reliability under normal operation.

An estimation of the technology readiness level (TRL) will be given according to the definitions given in D3.1, chapter 1.3. TRL will span from level 4 to possibly level 9 with the meaning hereafter reported:

- 4 Component and/or breadboard validation in laboratory environment
- 5 Component and/or breadboard validation in relevant environment
- 6 System/subsystem model or prototype demonstration in a relevant environment (ground or space)
- 7 System prototype demonstration in a space environment
- 8 Actual system completed and “flight qualified” through test and demonstration (ground or space)
- 9 Actual system “flight proven” through successful mission operations

2.1.4.3.6 Security

We do not evaluate this characteristic as it is deemed necessary and valid only for more complex systems.

2.1.4.3.7 Maintainability

Degree of effectiveness and efficiency with which a product or system can be modified by the intended maintainers.

Measurement function: $M_a = L$ where

M_a = Maintainability

L = Modifiability

Interpretation of test results: M_a value closer to 1 is better

Modifiability: degree to which a product or system can be effectively and efficiently modified without introducing defects or degrading existing product quality.

A score will be given according to the formula

$$K = \frac{A + B}{2}; \quad A, B \in \{0, 0.5, 1\}$$

where:

“A” takes into consideration if the tool is open-source with a score value of 1 or not with a score of 0. If the tool is opensource but the code is not clean and badly documented the score is 0.5.

“B” takes into account if the software tool is well supported either by an active community or in case of proprietary software by a quick and effective customer care service for bug fixing and user required customizations.

2.1.4.3.8 Portability

Degree of effectiveness and efficiency with which a system, product or component can be transferred from one hardware, software or other operational or usage environment to another.

Measurement function: $P_o = X$ where:

P_o = Portability

X = *Installability*

Interpretation of test results: P_o value closer to 1 is better

Installability: degree of effectiveness and efficiency with which a product or system can be successfully installed and/or uninstalled in a specified environment.

$$K = \frac{A + B}{2}; \quad A, B \in \{0, 0.5, 1\}$$

where:

“A” takes into consideration the way of installing. If there is an automatic installer the score is 1, the score is 0.5 if there is not an installer but a clear and effective installation procedure is available, 0 if not an installer nor a clear installation procedure is present.

“B” takes into consideration if the installer or the single steps of the installation procedure correctly alert about encountered problems and provide solutions e.g. “error: missing dependencies, please install third party software xyz”.

2.1.5 Preservation Platforms Assessment Criteria

In the following we describe the criteria and functions through which several digital preservation platforms are evaluated. As far as the assessment reference models are concerned, [ISO/IEC 25010, 2011] is the adopted standard. For the possible assessment measures the reference standard is [ISO/IEC 25023, 2012]. During year one assessment we evaluated two preservation systems: P4 and Archivematica (see D3.2). For the second year assessment we evaluated the version of Archivematica, DSpace and RODA (Fedora Commons). During the assessment of the digital preservation platforms we did not consider the rights management provided by these systems. This feature will be further discussed in D3.4.

2.1.5.1 Definition of Functions

In this section we list the functions used for the evaluation of Digital Preservation Platforms RO. The underlying assumption for the evaluated platforms is to be OAIS compliant, therefore each of the following functions is associated with one or more corresponding OAIS functional entities.

The level of requirement (mandatory, recommended, desirable) and a short description of each assessment function is also provided. The list below is not exhaustive because evaluating a platform is a complex activity and additional criteria could be defined. As explained in Section 2.5.1 of D3.2, we decided to focus on high level aspects which provide a clear understanding of the level or maturity (the TRL) of the solution.

Function	OAIS functional entity	Level of requirement	Description
M1 - GUI ingestion	Ingest	Mandatory	Ingestion using guided procedure offered by the GUI
M2 - Preservation of original content properties	Ingest	Mandatory	The original file received by the producer is stored in the archive
M3 - Support for AV formats	Ingest, Data Management	Mandatory	Support for AV formats selected for the Presto4U dataset
M4 - Preservation Workflows Management	Ingest, Preservation Planning	Mandatory	The platform implements workflows including tasks for content curation
M5 - Export of DIP	Access	Mandatory	Allow creation of Dissemination Packages for access
M6 - Periodic integrity checks of the material and storing information in the AIP	Preservation Planning	Mandatory	Periodic checks for file corruption (related also to availability of multiple copies for restore)
M7 - Format migration	Data Management	Mandatory	when format is at risk of obsolescence (a few tools working on it)
M8 - Ability to deal with large files	Archival Storage	Mandatory	Integrate storage technologies suitable even for huge files, for example larger than 10 GB (=20 min of MXF/D10)

M9 - Content quality control	Data Management	Mandatory	Integrate tools for QC
M10 - Virus check	Ingest, Data Management	Mandatory	Integrate tools for virus check of ingested content
R1 - Batch ingestion	Ingest	Recommended	Capability to ingest list of SIP files from CLI, managing ingestion queue
R2 - Support for METS	Ingest, Data Management, Access	Recommended	METS is used as a wrapper for SIP, AIP, DIP
R3- Support for PREMIS	Data Management	Recommended	PREMIS is used for preservation metadata and for logging preservation events
R4 - definition of requirements for restitution/playback	Access	Recommended	support reconstruction of the desired characteristics of the playback environment
R5 - Extension with Add-ons and plugins	All	Recommended	Integration of tools and services for specific purposes
R6 - Usage Documentation	All	Recommended	For archive administrators
R5 - Dashboard for job monitoring	All	Recommended	Provide real time information about active jobs (e.g. ingestion queue, periodic preservation tasks, ...), including used resources and status
R6- Automatic extraction of technical metadata	Data Management	Recommended	Extraction of technical metadata during ingestion
R7 - User profiles and ACL	Administration	Recommended	Manage user authentication and authorization, enable functionalities in the GUI according to permissions, etc
R8 - Creation of proxy copies (browsing quality)	Access	Recommended	Creation of low quality copy
R9- Multiple copies for redundancy	Ingest, Preservation Planning	Recommended	Ability to create device independent AIPs to ensure future access
D1 - Customize existing workflows	Ingest, Preservation Planning	Desirable	Allow configuration and customization of existing preservation tasks
D2 – Export of DIP to different formats	Access	Desirable	transcoding to format on Consumer's request
D3 - Export of AV content fragments	Access	Desirable	Partial restore
D4 – Ability to integrate with alternative collection management systems	Archival Storage	Desirable	Possibility to integrate with an alternative system providing functions different from preservation (e.g. cataloguing and searching)
D5 – Populate and draw data and statistics from collection management systems	Administration, Access	Desirable	Provide information about use of resources, number of accessed contents, etc

Table 8: Definition of Functions - Preservation and Platform Systems

2.1.5.2 Measurement Plan

Here below, for the Digital Preservation Platforms RO, the measurement plan is reported. We evaluate several features including the user interface, adopted technologies and usability. Before proceeding with the description of the features that will be assessed, it is worth noting that the approach adopted was to consider each platform as a black-box, focusing on input and output formats and interfaces. A further assessment could be

performed by taking into account every component of the platforms but such an evaluation goes beyond the aim of the project.

2.1.5.2.1 **Functional Suitability**

Degree to which a product or system provides functions that meet stated and implied needs when used under specified conditions.

Measurement function: $FS = (X+Z)/2$ where

FS = Functional Suitability

X = Functional Completeness

Z = Functional Appropriateness

X indicates how complete is the implementation according to requirement specifications.

As reported in Section 2.1.1 of D3.2, X can be calculated as follows:

$$X = (X1 + X2 * 0.5 + X3 * 0.25) / 1.75$$

Where $X1 = 1 - (A/B)$, $X2 = 1 - (C/D)$ and $X3 = 1 - (E/F)$. A is the number of missing or unsatisfying mandatory functions, B is the number of mandatory functions assessed in the evaluation, C is the number of missing or unsatisfying recommended functions assessed in the evaluation, D is the number of recommended functions, E is the number of missing or unsatisfying desirable functions assessed in the evaluation and F is the number of desirable functions.

Z describes how many functions with no problems are implemented for the appropriate functions for pursuing a specific task. Z can be calculated as:

$$Z = A/B$$

Where A is the sum of the scores of the implemented functions and B is the total amount of implemented functions.

Interpretation of test results: FS value closer to 1 is better. The list of functions for the preservation platform considered during the assessment is presented in Section 2.5.1 of D3.2.

2.1.5.2.2 **Performance Efficiency**

Measurement function: $PE = Z$ where

$PE = \text{Performance Efficiency}$

$Z = \text{Capacity}$

Useful element for the evaluation of the capacity can be: the number of requests or simultaneous access per unit of time; the number of simultaneous jobs accepted in the ingestion queue or the number of tasks executed in parallel during a preservation workflow.

Such elements are strictly related to the hardware of the system into which the platforms are executed. For instance, since it is common for a new job or online request to throw a new thread, the availability of several computational units would improve the operation time of the platforms.

Due to the previous considerations, if the platform architecture allows a uniform distribution of the tasks, the capacity is scalable and thus the platform should get a good evaluation.

Interpretation of test results: PE closer to 1 is better.

2.1.5.2.3 **Compatibility**

Measurement function: $Co = (X+Y)/2$ where

$Co = \text{Compatibility}$

$X = \text{Co-existence}$

$Y = \text{Interoperability}$

As explained in Section 2.1.3 of D3.2, X indicates how flexible is the product in sharing its environment with other products without adverse impacts on other products.

It is possible to evaluate if the platform requires an exclusive usage of a component such as the database. In case the database can be shared among other systems, the platform should get a good score for this feature (between 0 and 1).

Y indicates how accurately is implementation of data exchange format determined between linking systems. It can be expressed as:

$Y = A/B$

Where A is the number of formats into which data can be exported in order to be exchanged with other platforms. B is the total number of data exportation formats provided by the platforms being assessed.

Interpretation of test results: Co value closer to 1 is better.

Note: This measurement is quite important for preservation systems because demonstrates also the level of integration for different technologies and systems used by the platform to implement the OAIS model. Possible measure can include the use of external systems for storage only or for complete collection management, taking into account the complexity of the integration, the interfacing mechanism and any known limitation for example in terms of supported protocols or technologies.

2.1.5.2.4 Usability

Degree to which a product or system can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.

Measurement function: $Us = (K+L)/2$ where

$Us =$ **Usability**

$K =$ Operability

$L =$ User error protection

According to its definition, the operability indicates the degree to which the platform has attributes that make it easy to operate and control. A good estimation of K may come from the evaluation of the user interface provided by the platform. In case a clear and intuitive interface is provided the platform should get a good mark (between 0 and 1).

L describes how many functions have incorrect operation avoidance capabilities. This feature can be regarded as the degree to which the platform prevents the users from making mistakes, especially during the ingest process, that could affect the preservation of data. In particular it can be evaluated as:

$$L = (A+B+C+D+E)/5$$

Where A indicates whether there are required field to fill during the ingest process in order to clearly identify the data being ingested. B indicates if the platform checks the input formats to determine if they are compatible with its preservation capabilities (for instance the platform must be capable of migrating the format to another one). C indicates whether a check of the metadata is performed. D is the degree to which the user is guided through

the ingestion process and E indicates if a check of the authenticity of the data is performed.

Interpretation of test results: Us value closer to 1 is better.

Note: for preservation systems the operability can be mainly associated to the user interface, because it should provide user all required information to interact with the platform, although with different levels (a basic user should be able to perform a limited number of operations with respect to an administrator, which should be ready to perform complex operations to solve problems). The user error protection can also be associated to the user interface, but should mainly reflect the capability of the system to prevent wrong operations which can have disrupting consequences (e.g. deletion of content or execution of wrong resource consuming tasks).

2.1.5.2.5 Reliability

Degree to which a system, product or component performs specified functions under specified conditions for a specified period of time.

Measurement function: $Re = (H+J+K+L)/4$ where

Re = Reliability

H = Maturity

J = Availability

K = Fault Tolerance

L = Recoverability

As far as H is concerned, since the platforms taken into account are developed, supported and adopted by communities of users, this value should give a qualitative estimation of how wide the community behind the platform is and its degree of adoption. A score between 0 and 1 will be assigned.

J represents the availability of the platform. Since each of these systems is based upon web services, it is possible to assign a mark between 0 and 1 according to how the web services can be monitored by the user.

K concerns how the platform can deal with user's errors or other failures without compromising the whole operation. It can be defined as:

$$K = (A+B+C)/3$$

Where A indicates if the platform allows to save a complete backup in order to restore the overall state of the platform itself in case of failure. B indicates the degree to which making a mistake affect the normal operability of the system. C indicates if the platform provides a validation mechanism for the ingestion process.

L indicates what is (the average) time the system takes to complete recovery from a failure. It is possible to take into account a given task, such as the ingestion process, and evaluate how the system reacts to the occurrence of a failure. In case the platform allows the user to cope with the failure and continue the ingestion the recoverability value should be close to 1. If, on the other hand, the platform requires the user to start the ingestion process from the beginning, this value should be close to 0.

Interpretation of test results: Re value closer to 1 is better.

2.1.5.2.6 Security

Degree to which a product or system protects information and data so that persons or other products or systems have the degree of data access appropriate to their types and levels of authorization.

Measurement function: $Se = (H+J+K+L+M)/5$ where

Se = Security

H = Confidentiality

J = Integrity

K = Non-repudiation

L = Accountability

M = Authenticity

According to Section 2.1.6 of D3.2, H, J, K, L and M can be defined as follows:

H indicates how controllable is the access to the system. Since the platforms take advantage of web services to manage the ingested data, the security level provided by these web services is related to the degree of confidentiality.

J describes to what extent the system prevents unauthorised access to the data. This feature is closely related to the previous one so the security of the web services has to be taken into account.

K indicates what proportion of events requiring non-repudiation are processed. In order to satisfy this requirement the platform must be able to prove that an action has been performed so that it cannot be repudiated later. In case the system is provided with this capability it should get a high mark (from 0 to 1).

L describes how complete is the audit trail concerning the user access to the system and data. For the kind of systems being assessed, this feature may be related to the ACL capability so that the platform can assign a different access level to administrators with respect to users. The more complete is the set of rules that can be established, the higher is the score (between 0 and 1).

M indicates how well does the system authenticate the identity of a subject or resource. It is implemented as:

$$M = A/B$$

Where A is the number of provided authentication methods (e.g., ID/password or IC card) and B is the total number of authentication methods specified in the requirements (e.g., ID/password or IC card).

Note: Confidentiality and integrity are often based on user authorization and authentication, with the definition of appropriate ACLs and mechanisms for protecting data from unauthorized access.

Interpretation of test results: Se value closer to 1 is better.

2.1.5.2.7 Maintainability

Degree of effectiveness and efficiency with which a product or system can be modified by the intended maintainers.

Measurement function: $M_a = (H+K+L+M)/4$ where

$M_a =$ **Maintainability**

H = Modularity

K = Analysability

L = Modifiability

M = Testability

In Section 2.1.7 of D3.2, H, K, L and M are described as follows.

H measures how strong is the relation between the components in a system or computer program. Certainly the platforms being assessed are made up of several components that have to interact with each other in order to make the system work properly. Considering the large communities of users and developers supporting these platforms, the interaction of the various components is granted by the maturity of the systems. Therefore an element that can be taken into account for the assessment is the possibility for the user to store data into a cloud storage. Keeping data separated from the system can be a benefit in case of local failures.

K indicates whether users can easily identify specific operation which caused failures. It is possible to consider the ingest process where the most part of errors may occur. In case the platform warns the user about failures and indicates the task that caused it, then the system should get a good mark (between 0 and 1).

L indicates if the maintainer can easily modify the software to meet some modification requirement. An example of whether this requirement is satisfied is the possibility to switch from one database to another. This feature is related to the modularity.

M describes how completely are test functions and facilities implemented. It can be calculated as follows:

$$M = (A+B+C)/3$$

Where A is 1 in case the platform allows the user to perform dry run in order to verify the correctness of the operation, B is 1 if the platform provides diagnostic tools within its user interface and C is one in case it is possible to run a demo version of the platform in order to perform tests without compromising the actual data.

Interpretation of test results: Ma value closer to 1 is better.

2.1.5.2.8 Portability

Degree of effectiveness and efficiency with which a system, product or component can be transferred from one hardware, software or other operational or usage environment to another.

Measurement function: $P_o = (X+Y+Z)/3$ where

P_o = **Portability**

X = Adaptability

Y = Installability

Z = Replaceability

The description of X, Y and Z is reported in Section 2.1.8 of D3.2. The evaluation of these features may differ from the one described in Section 2.1.8 in order to better adapt to the assessment of digital platforms.

X indicates whether the software system is capable enough to adapt itself to different hardware environment. It is calculated as:

$$X = 1 - (A/B)$$

Where A is the number of operational functions of which tasks were not completed or not enough resulted to meet adequate levels during testing and B is the total number of functions which were tested in different hardware environment.

Y gives an idea of how much time and trouble is required to make an install. As far as this feature is concerned, the platform will be evaluated according to how clearly and completely is the installation process described in the documentation.

Z measures the degree to which the system can be replaced by another one with the same purpose. The adoption of standard is a relevant element for the evaluation of this feature. Another element to take into account is whether is possible for the platform to be integrated with another one.

Interpretation of test results: Po value closer to 1 is better.

Note: the evaluation of these sub characteristics is affected by: the dependencies to be taken into account during migration from one environment to the other, the requirements to be satisfied before installing the platform, the possibility to replace the platform with a similar one with additional features without changing the whole environment.

3 Results of Research Outputs Assessment – Year 2

This chapter will present the detailed quantitative evaluation results of the tools chosen for assessment in year 2.

3.1 Metadata mapping

3.1.1 Assessment results for Metadata Interoperability (MINT) toolset for EBUCore

In the following, the evaluation of the Metadata Interoperability (MINT) toolset for EBUCore done according to the measurement plan defined in Section 2.1.1 is presented.

The following reference system has been used: Intel Core i7-3770, 3.4 GHz, 8 GB RAM, Intel HD Graphics 4000

3.1.1.1 Functional suitability

Degree to which a product or system provides functions that meet stated and implied needs when used under specified conditions.

Measurement function: $FS = (X + Y + Z) / 3$ (where X, Y, Z are the scores computed as in the following)

FS = Functional Suitability

X = Functional Completeness

Y = Functional Correctness

Z = Functional Appropriateness

In particular the ability to reach a defined goal using the tool can be considered

Interpretation of test results: FS value closer to 1 is better

Functional Completeness:

1. Measure: functional metadata formats coverage
2. Description of measure: Evaluation of the metadata formats supported and the constructs of data model supported. Measured by comparing against widely adopted metadata formats/models.
3. Measurement function: $X = 1/n \sum_{i=1..n} w_i f_i$, with n being the number the functions, w_i the weight of the function (mandatory = 1, recommended = 0.75, desired = 0.5) and f_i the degree to which the function is provided [0..1] (with 1 = completely provided).

Table 9 lists the functional completeness of MINT.

Functions	Levels of Need	Description	Functional Completeness
<i>Metadata input formats</i>		Support for a metadata format as source format of the mapping process	

Functions	Levels of Need	Description	Functional Completeness
Dublin Core	Mandatory		1
ESE	Desirable		1
EDM	Desirable		0
EBU Core	Mandatory (for broadcasting related CoPs) / Recommended		1
MPEG-7	Recommended (for CoPs using automatic content analysis)		1
LIDO	Mandatory (for museum/gallery related) / Recommended		1
EAD	Mandatory (for non-a/v archive related) / Recommended		1
<i>Metadata output formats</i>		Support for a metadata format as target format of the mapping process	
Dublin Core	Mandatory		0
ESE	Recommended		0
EDM	Recommended		0
EBU Core	Mandatory (for broadcasting related CoPs) / Recommended		1
MPEG-7	Recommended (for CoPs using automatic content analysis)		0
LIDO	Mandatory (for museum/gallery related) / Recommended		0
EAD	Mandatory (for non-a/v archive related) / Recommended		0
option to add custom formats	Recommended	Support for adding new metadata formats	0.5
XML representation support	Mandatory	Support for metadata documents in XML format	1
RDF representation support	Recommended	Support for metadata documents in RDF format	0
<i>Metadata model constructs</i>			
single -> multiple elements	Recommended	Support mapping a single element into a set of elements (e.g., string into structured)	1
multiple -> single elements	Mandatory	Support mapping a set of elements into a single elements (e.g., structured into string)	1
structure using context elements	Mandatory	Define mapping of content structure constructs using contextual elements	1
conditional mapping based on element/attribute	Mandatory	Define mapping rules that are conditioned on values of elements or attributes	1

Functions	Levels of Need	Description	Functional Completeness
values			
map collections	Recommended	Support for mapping of collections of metadata records rather than single metadata records only	1
merge string values	Mandatory	Support merging values of separate string values into one string value	1
split string values	Recommended	Support splitting a string value into a set of separate string values	1
number of levels in data structure	Mandatory: 2 / Recommended: 2+	Support for number of hierarchy levels in the document structure	1
start from example(s)	Recommended	Support initiating mapping from example documents	1
start from schema	Recommended	Support initiating mapping from a schema instance	0
configuration user interface	Mandatory	Provision of a configuration user interface (instead of/in addition to configuration files)	1
<i>user interface</i>			
drag & drop mappings	Recommended	Support of drag&drop for configuring the mappings	1
preview	Mandatory	Provide preview of configured mappings	1
map constructs not found in available examples	Recommended	Support the definition of mappings for constructs not found in one of the examples	0

Table 9: Functional completeness of MINT.

X = Functional Completeness = 0.66

Functional Correctness:

1. Measure: functional correctness of mapping
2. Description of the measure: Correctness and completeness of the mapping between a pair of formats.
Note: The evaluation of the mapping can be performed and validated by an expert.
3. Measurement function: $Y = 1/m \sum_{i=1..m} c_i$, with m being the number of mapping checked, and c_i the assessment of the correctness of the mapping (with 1 = fully correct).

Table 10 lists the functional correctness of MINT.

<i>Metadata input format</i>	<i>Metadata output format</i>	Functional Completeness
Dublin Core	MPEG-7	0
MPEG-7	Dublin Core	0
EBU Core	Dublin Core	0
EBU Core	MPEG-7	0
Dublin Core	EAD	0
Dublin Core	EBU Core	1
MPEG-7	EBU Core	1
Dublin Core	EDM	0

<i>Metadata input format</i>	<i>Metadata output format</i>	Functional Completeness
Dublin Core	MPEG-7	0

Table 10: Functional correctness of MINT.

Y = Functional Correctness = 0.22

Functional Appropriateness:

1. Measure: functional appropriateness of the mapping tools
2. Description: The evaluation considers mappings required in common preservation workflow steps (ingest, B2B exchange, export to Europeana) as assessed by an expert.
3. Measurement function: $Z = 1/p \sum_{i=1..p} a_i$, with p being the number of workflows checked, and a_i the assessment of the appropriateness for the workflow (with 1 = fully appropriate).

Table 11 lists the functional appropriateness of MINT.

<i>Workflow</i>	Functional Appropriateness
Ingest	0.6
B2B exchange	0.5
Export to Europeana	0

Table 11: Functional appropriateness of MINT

Z = Functional Appropriateness = 0.36

FS = Functional Suitability = $(X+Y+Z) / 3 = (0.66 + 0.22 + 0.36) / 3 = 0.41$

3.1.1.2 Performance efficiency

Performance relative to the amount of resources used under stated conditions.

Measurement function: $PE = (X+Y+Z) / 3$ where X,Y,Z are the scores computed as in the following:

PE = Performance Efficiency
 X = Time behaviour
 Y = Resource utilization
 Z = Capacity

Interpretation of test results: PE smaller is better

Time behaviour:

Measure: The mean processing time X in milliseconds for performing a defined set of mapping problems.

Interpretation of test results: X varies from 0 to infinite. Smaller is better.

Time for executing a mapping including 100 metadata elements.

Maximum time: 500 ms

$X = \text{Time behaviour} = 300 \text{ ms} \rightarrow 0.6$

Resource utilization:

Degree to which the amounts and types of resources used by a product or system when performing its functions meet requirements

1. Measure: (Mean) CPU/RAM utilization
2. Description of the measure: How much CPU time/RAM is used to perform a given task
3. Measurement function:
 $Y = 0.5 * (\text{fraction of CPU} + \text{fraction of RAM})$ actually used to perform a task on a reference system
4. Interpretation of test results: Y smaller is better.

$Y = 0.5 * (0.2 + 0.2) = 0.2$

Capacity:

Degree to which the maximum limits of a product or system parameter meet requirements

Measure: Maximum throughput using a specific reference configuration

1. Measurement function:
 $Z = A/B$ where
 $A = \text{operation time}$
 $B = \text{the total no. of processed documents}$
2. Interpretation of test results: Z smaller is better.

$Z_{\max} = 0.60 \text{ min / document (score 1.0)}$

$Z = A / B = 1 \text{ min} / 3 = 0.33 \rightarrow \text{normalised score } 0.56$

$PE = (X+Y+Z) / 3 = (0.6 + 0.2 + 0.56) / 3 = 0.45$

3.1.1.3 Compatibility

Degree to which a product, system or component can exchange information with other products, systems or components, and/or perform its required functions, while sharing the same hardware or software environment.

Measurement function: $Co = Y$ where

Co = Compatibility
 Y = Interoperability

Interpretation of test results: Co value larger is better

Compatibility (Co) = $Y = 0.33$

Interoperability:

Degree to which two or more systems, products or components can exchange information and use the information that has been exchanged

1. Measure: Supported service interfaces for information exchange (metadata files, REST, SOAP)
2. Description of the measure: the service interfaces are smoothly exchanged with other software or systems
3. Measurement function:
 $Y = A / B$ where
 A = number of interfaces for information exchange
 B = total number of interfaces to be supported
4. Interpretation of test results: Y varies from 0 to infinite. Usually, larger is better.

Interoperability = $Y = (1 + 0 \text{ (REST)} + 0 \text{ (SOAP)}) / 3 = 0.33$

Compatibility = $Co = Y = 0.33$

3.1.1.4 Usability

Degree to which a product or system can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.

Measurement function: $Us = N$, where

Us = Usability
 N = Accessibility

Interpretation of test results: Us value closer to 1 is better

Accessibility:

1. Measure: The control elements of the user interface that provide accessibility by adaptable text sizes, alternative image texts and supporting multiple input devices.

2. Measurement function:

$$N = A / B \text{ where}$$

A = number of control elements supporting accessibility options

B = total number of control elements checked

3. Interpretation of test results: N varies from 0 to 1. Larger is better.

$$A = 3$$

$$B = 10$$

$$\text{Accessibility} = N = 3 / 10 = 0.3$$

$$\text{Usability} = U_s = N = 0.3$$

3.1.1.5 Reliability

Degree to which a system, product or component performs specified functions under specified conditions for a specified period of time.

Measurement function: $Re = L$, where

Re = Reliability

L = Recoverability

Interpretation of test results: Re value closer to 1 is better

Recoverability:

The possibility of using the tool in the following situations:

- incorrect input file format
- load incomplete mapping definition
- user creates contradicting mapping
- wrong drag & drop operation
- example file in wrong format

$L = 1/n \sum_{i=1..n} r_i$, where n is the number of situations tested, and r_i is the assessment of the recoverability in that situation [0..1] (1 = fully recoverable without side effects).

$$\text{Recoverability} = L = (1 + 1 + 1 + 1 + 0 \text{ (example file in wrong format)}) / 5 = 0.8$$

$$\text{Reliability} = Re = L = 0.8$$

3.1.1.6 Maintainability

Degree of effectiveness and efficiency with which a product or system can be modified by the intended maintainers.

Measurement function: $Ma = (H+J) / 2$ where

Ma = Maintainability

H = Modularity

J = Reusability

Interpretation of test results: Ma value closer to 1 is better

Modularity:

1. Measure: Easiness of adding new formats, data types and mapping rules.
2. Measurement function:
H = assessment of effort for adding a new mapping
3. Interpretation of test results: H varies from 0 to 1. Closer to 1 is better.

Modularity = H = 0.7

Reusability:

Degree to which an asset can be used in more than one system, or in building other assets

1. Measure: Integration in other system based on documented interfaces and exchange formats.
2. Description of the measure: Supported service interfaces.
3. Measurement function:

$$J = A / B$$

A = number of service interfaces/exchange formats supported

B = total number of service interfaces/exchange formats considered

Interpretation of test results: J varies from 0 to 1. Usually, closer to 1 is better.

A = 1

B = 2

Reusability = $J = 1 / 2 = 0.5$

Maintainability = $(H+J) / 2 = (0.7 + 0.5) / 2 = 0.6$

3.1.2 Assessment results for PrestoPRIME Metadata Mapping Tool

In the following, the evaluation of the PrestoPRIME Metadata Mapping Tool done according to the measurement plan described in Section XX is presented.

The following reference system has been used: Intel Core i7-3770, 3.4 GHz, 8 GB RAM, Intel HD Graphics 4000.

3.1.2.1 Functional suitability

Degree to which a product or system provides functions that meet stated and implied needs when used under specified conditions.

Measurement function: $FS = (X+Y+Z) / 3$ (where X,Y,Z are the scores computed as in the following)

FS = Functional Suitability

X = Functional Completeness

Y = Functional Correctness

Z = Functional Appropriateness

In particular the ability to reach a defined goal using the tool can be considered

Interpretation of test results: FS value closer to 1 is better

Functional Completeness:

1. Measure: functional metadata formats coverage
2. Description of measure: Evaluation of the metadata formats supported and the constructs of data model supported. Measured by comparing against widely adopted metadata formats/models.
3. Measurement function: $X = 1/n \sum_{i=1..n} w_i f_i$, with n being the number the functions, w_i the weight of the function (mandatory = 1, recommended = 0.75, desired = 0.5) and f_i the degree to which the function is provided [0..1] (with 1 = completely provided).

Table 12 lists the functional completeness of PrestoPRIME Metadata Mapping Tool.

Functions	Levels of Need	Description	Functional Completeness
<i>Metadata input formats</i>		Support for a metadata format as source format of the mapping process	
Dublin Core	Mandatory		1
ESE	Desirable		0
EDM	Desirable		0
EBU Core	Mandatory (for broadcasting related CoPs) / Recommended		0,5
MPEG-7	Recommended (for CoPs using automatic content analysis)		1

Functions	Levels of Need	Description	Functional Completeness
LIDO	Mandatory (for museum/gallery related) / Recommended		0
EAD	Mandatory (for non-a/v archive related) / Recommended		0.5
<i>Metadata output formats</i>		Support for a metadata format as target format of the mapping process	
Dublin Core	Mandatory		1
ESE	Recommended		0
EDM	Recommended		0.5
EBU Core	Mandatory (for broadcasting related CoPs) / Recommended		0
MPEG-7	Recommended (for CoPs using automatic content analysis)		0.5
LIDO	Mandatory (for museum/gallery related) / Recommended		0
EAD	Mandatory (for non-a/v archive related) / Recommended		0.5
option to add custom formats	Recommended	Support for adding new metadata formats	0.5
XML representation support	Mandatory	Support for metadata documents in XML format	1
RDF representation support	Recommended	Support for metadata documents in RDF format	0.5
<i>Metadata model constructs</i>			
single -> multiple elements	Recommended	Support mapping a single element into a set of elements (e.g., string into structured)	1
multiple -> single elements	Mandatory	Support mapping a set of elements into a single elements (e.g., structured into string)	1
structure using context elements	Mandatory	Define mapping of content structure constructs using contextual elements	1
conditional mapping based on element/attribute values	Mandatory	Define mapping rules that are conditioned on values of elements or attributes	0.4
map collections	Recommended	Support for mapping of collections of metadata records rather than single metadata records only	0.5
merge string values	Mandatory	Support merging values of separate string values into one string value	1
split string values	Recommended	Support splitting a string value into a set of separate string values	1
number of levels in data structure	Mandatory: 2 / Recommended: 2+	Support for number of hierarchy levels in the document structure	1
start from example(s)	Recommended	Support initiating mapping from example documents	0
start from schema	Recommended	Support initiating mapping from a schema instance	1

Functions	Levels of Need	Description	Functional Completeness
configuration user interface	Mandatory	Provision of a configuration user interface (instead of/in addition to configuration files)	1
<i>user interface</i>			
drag & drop mappings	Recommended	Support of drag&drop for configuring the mappings	1
preview	Mandatory	Provide preview of configured mappings	1
map constructs not found in available examples	Recommended	Support the definition of mappings for constructs not found in one of the examples	

Table 12: Functional completeness of PrestoPRIME Metadata Mapping Tool.

X = Functional Completeness = 0.63

Functional Correctness:

1. Measure: functional correctness of mapping
2. Description of the measure: Correctness and completeness of the mapping between a pair of formats.
Note: The evaluation of the mapping can be performed and validated by an expert.
3. Measurement function: $Y = 1/m \sum_{i=1..m} C_i$, with m being the number of mapping checked, and c_i the assessment of the correctness of the mapping (with 1 = fully correct).

Table 13 lists the functional correctness of PrestoPRIME Metadata Mapping Tool.

<i>Metadata input format</i>	<i>Metadata output format</i>	<i>Functional Completeness</i>
Dublin Core	MPEG-7	1
MPEG-7	Dublin Core	1
EBU Core	Dublin Core	0.3
EBU Core	MPEG-7	0.3
Dublin Core	EAD	0.8
Dublin Core	EBU Core	0
MPEG-7	EBU Core	0
Dublin Core	EDM	0.8
Dublin Core	MPEG-7	1

Table 13: Functional correctness of PrestoPRIME Metadata Mapping Tool.

Y = Functional Correctness = 0.57

Functional Appropriateness:

1. Measure: functional appropriateness of the mapping tools
2. Description: The evaluation considers mappings required in common preservation workflow steps (ingest, B2B exchange, export to Europeana) as assessed by an expert.

3. Measurement function: $Z = 1/p \sum_{i=1..p} a_i$, with p being the number of workflows checked, and a_i the assessment of the appropriateness for the workflow (with 1 = fully appropriate).

Table 14 lists the functional appropriateness of PrestoPRIME Metadata Mapping Tool.

<i>Workflow</i>	Functional Appropriateness
Ingest	0.5
B2B exchange	0.5
Export to Europeana	0.8

Table 14: Functional appropriateness of PrestoPRIME Metadata Mapping Tool.

$Z = \text{Functional Appropriateness} = 0.6$

$FS = \text{Functional Suitability} = (X+Y+Z) / 3 = (0.63 + 0.57 + 0.6) / 3 = 0.6$

3.1.2.2 Performance efficiency

Performance relative to the amount of resources used under stated conditions.

Measurement function: $PE = (X+Y+Z) / 3$ where X, Y, Z are the scores computed as in the following:

PE = Performance Efficiency
 X = Time behaviour
 Y = Resource utilization
 Z = Capacity

Interpretation of test results: PE smaller is better

Time behaviour:

Measure: The mean processing time X in milliseconds for performing a defined set of mapping problems.

Interpretation of test results: X varies from 0 to infinite. Smaller is better.

Time for executing a mapping including 100 metadata elements.

Maximum time: 500 ms

$X = \text{Time behaviour} = 450 \text{ ms} \rightarrow 0.9$

Resource utilization:

Degree to which the amounts and types of resources used by a product or system when performing its functions meet requirements

1. Measure: (Mean) CPU/RAM utilization

2. Description of the measure: How much CPU time/RAM is used to perform a given task
3. Measurement function:

$$Y = 0.5 * (\text{fraction of CPU} + \text{fraction of RAM}) \text{ actually used to perform a task on a reference system}$$
4. Interpretation of test results: Y smaller is better.

$$Y = 0.5 * (0.2 + 0.15) = 0.175$$

Capacity:

Degree to which the maximum limits of a product or system parameter meet requirements

Measure: Maximum throughput using a specific reference configuration

3. Measurement function:

$$Z = A/B \text{ where}$$
 - A = operation time
 - B = the total no. of processed documents
4. Interpretation of test results: Z smaller is better.

$$Z_{\max} = 0.60 \text{ min / document (score 1.0)}$$

$$Z = A / B = 1 \text{ min} / 4 = 0.25 \rightarrow \text{normalised score } 0.42$$

$$PE = (X+Y+Z) / 3 = (0.9 + 0.175 + 0.42) / 3 = 0.50$$

3.1.2.3 Compatibility

Degree to which a product, system or component can exchange information with other products, systems or components, and/or perform its required functions, while sharing the same hardware or software environment.

Measurement function: $Co = Y$ where

$Co = \text{Compatibility}$
 $Y = \text{Interoperability}$

Interpretation of test results: Co value larger is better

Interoperability:

Degree to which two or more systems, products or components can exchange information and use the information that has been exchanged

1. Measure: Supported service interfaces for information exchange (metadata files, REST, SOAP)

2. Description of the measure: the service interfaces are smoothly exchanged with other software or systems
3. Measurement function:
 $Y = A / B$ where
 $A =$ number of interfaces for information exchange
 $B =$ total number of interfaces to be supported
4. Interpretation of test results: Y varies from 0 to infinite. Usually, larger is better.

Interoperability = $Y = (1 + 1 + 0 \text{ (SOAP)}) / 3 = 0.66$

Compatibility = $Co = Y = 0.66$

3.1.2.4 Usability

Degree to which a product or system can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.

Measurement function: $Us = N$, where

$Us =$ Usability

$N =$ Accessibility

Interpretation of test results: Us value closer to 1 is better

Accessibility:

1. Measure: The control elements of the user interface that provide accessibility by adaptable text sizes, alternative image texts and supporting multiple input devices.
2. Measurement function:
 $N = A / B$ where
 $A =$ number of control elements supporting accessibility options
 $B =$ total number of control elements checked
3. Interpretation of test results: N varies from 0 to 1. Larger is better.

$A = 3$

$B = 10$

Accessibility = $N = 3 / 10 = 0.3$

Usability = $Us = N = 0.3$

3.1.2.5 Reliability

Degree to which a system, product or component performs specified functions under specified conditions for a specified period of time.

Measurement function: $Re = L$, where

Re = Reliability

L = Recoverability

Interpretation of test results: Re value closer to 1 is better

Recoverability:

The possibility of using the tool in the following situations:

- incorrect input file format
- load incomplete mapping definition
- user creates contradicting mapping
- wrong drag & drop operation
- example file in wrong format

$L = 1/n \sum_{i=1..n} r_i$, where n is the number of situations tested, and r_i is the assessment of the recoverability in that situation [0..1] (1 = fully recoverable without side effects).

Recoverability = L = $(1 + 0 \text{ (incomplete mapping definition)} + 1 + 1 + 1) / 5 = 0.8$

Reliability = Re = L = 0.8

3.1.2.6 Maintainability

Degree of effectiveness and efficiency with which a product or system can be modified by the intended maintainers.

Measurement function: $Ma = (H+J) / 2$ where

Ma = Maintainability

H = Modularity

J = Reusability

Interpretation of test results: Ma value closer to 1 is better

Modularity:

1. Measure: Easiness of adding new formats, data types and mapping rules.

2. Measurement function:

H = assessment of effort for adding a new mapping

3. Interpretation of test results: H varies from 0 to 1. Closer to 1 is better.

$$\text{Modularity} = H = 0.4$$

Reusability:

Degree to which an asset can be used in more than one system, or in building other assets

1. Measure: Integration in other system based on documented interfaces and exchange formats.
2. Description of the measure: Supported service interfaces.
3. Measurement function:

$$J = A / B$$

A = number of service interfaces/exchange formats supported

B = total number of service interfaces/exchange formats considered

Interpretation of test results: J varies from 0 to 1. Usually, closer to 1 is better.

$$A = 2$$

$$B = 2$$

$$\text{Reusability} = J = 2 / 2 = 1$$

$$\text{Maintainability} = (H+J) / 2 = (0.4 + 1) / 2 = 0.7$$

3.2 Vocabulary mapping

3.2.1 Assessment results for Amalgame

The evaluation of the tool was foreseen in four phases:

1. Preparation of test vocabularies
2. Installation of Amalgame tool
3. Creation of alignment mechanism
4. Test mapping with data

3.2.1.1 Preparation of vocabularies

Given the engagement of JRS in various different cultural heritage domains a number of vocabularies could be used. Those vary in the aspects they describe but also in complexity (e.g. starting with materials of things over rather simple descriptions of geographic names and places up to complex descriptions of audio-visual content or related preservation information).

The tests were foreseen as a stepwise approach starting with “simple data” and increasing their degree of complexity over the steps.

Two kinds of test data were produced:

- **Geographic information** was extracted in-house at JR from a geographic thesaurus available in a relational database. These data were exported to SKOS formatted RDF files. Although geographic information can be thought of being structured in a rather simple way other difficulties may arise with such kinds of vocabularies (e.g. variations over time or due to different languages etc.). Therefore the exported data were made available in one file regarding Switzerland related information and another one dealing with information other places in the world. This distinction was made as the Swiss related data were especially used in use cases of a cultural heritage institution and were therefore handled with much more care than other places which were not of intensive use by that organization. For dealing with geographic data we tried to figure out other commonly used data sets that can be used for testing the mapping between vocabularies (i.e. the Getty Thesaurus of Geographic Names – TGN). We identified several partial data sets from this thesaurus referring to countries with German language (i.e. Austria, Germany, Liechtenstein and Switzerland).
- A second data set was around **preservation of av content**. We extracted data from a rich MPEG-7 data set and simplified data therein for our tests. It was intended to use the fully flavoured data set in a later test.

3.2.1.2 Installation of Amalgame tool

For the installation of the Amalgame tool we used a Windows PC / Server.

The software for the Amalgame tool is available at

<http://semanticweb.cs.vu.nl/amalgame>.

The software consists of a number of different components:

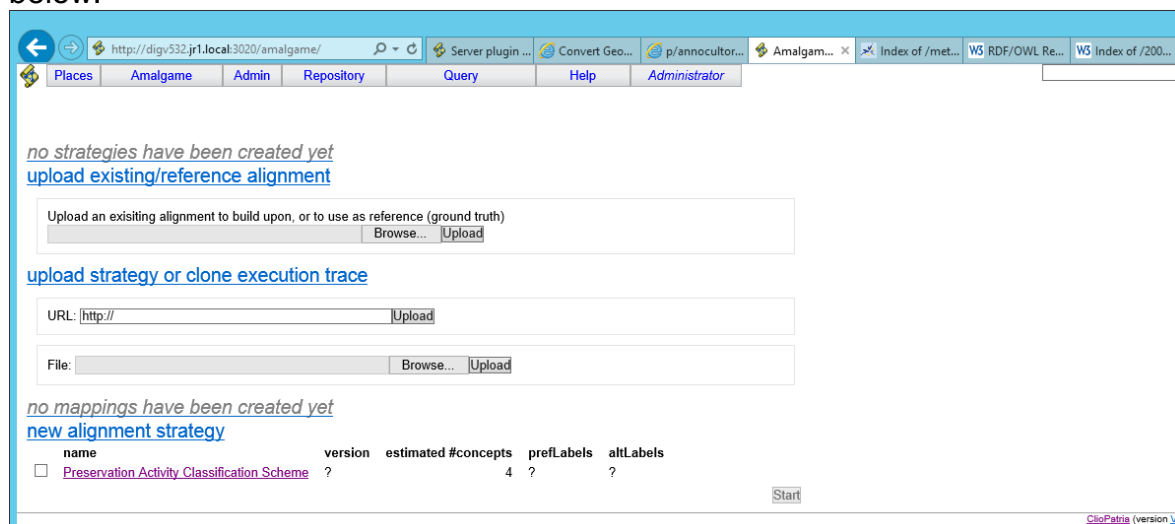
- The so called **SWI-Prolog** package which is made available via a GitHub repository was installed first. SWI-Prolog offers a comprehensive free Prolog environment and is widely used in research and education as well as commercial applications. The development versions can be found at <http://www.swi-prolog.org/download/devel>. In the first runs we tried with stable versions but realized that we would have to stick to newer versions as they support the necessary functions even if those versions may not be the current stable ones. Finally we took version 7.1.27.
- A basic framework called **ClioPatria** which is based on SWI-Prolog is necessary as well. ClioPatria provides an RDF application platform and can be found at <https://github.com/ClioPatria/ClioPatria>. Similar to SWI-Prolog it was announced that we should stick to the most recent version which was version V3.0.0-190-gba129d8 in our case. Unfortunately changing to newer versions of the framework was not fully transparent and caused inconsistencies in the software which made necessary to do total new installations.
- On top of the installation packages, the **Amalgame** application, were added to the ClioPatria framework. Due to the inconsistencies already created with the different versions of ClioPatria the package Amalgame did not perform correctly as well. Both the inconsistencies in ClioPatria and the malfunction Amalgame could not be observed easily as no kind of errors or appeared during installation or when running the software.

Technical difficulties during the installation process could be partially solved by contacting staff at Vrije Universiteit Amsterdam (VU) who have implemented the tools and are maintaining the source code.

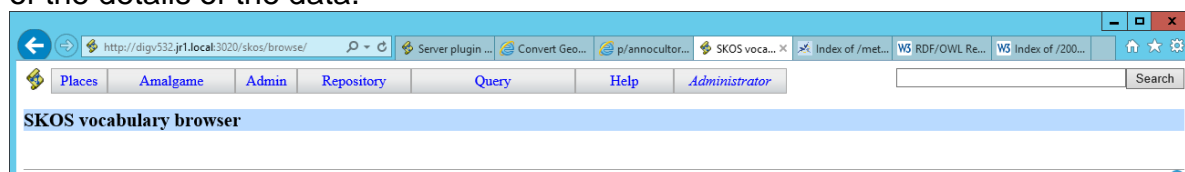
3.2.1.3 Creation of alignment mechanism

With the previously prepared test data sets and the installed software the tests were started. A number of issues appeared during the tests. Figuring them out was hard given that only very little documentation could be found about the Amalgame software and its use:

- Uploading a vocabulary worked fine and rather quick but the uploaded data – although found being correct (which was also confirmed by staff at VU) – did not seem to be correctly added to the repository. The software did not come up with correct information as shown for *version*, *prefLabels* and *altLabels* in the screenshot below.



- When trying to browse through an uploaded SKOS vocabulary (available through the vocabulary link in the previous screenshot) an empty screen was shown instead of the details of the data.



The developers were asked about this and previously mentioned issues. A potential reason for the various problems and bugs which appeared can be inconsistencies in the software due to transition from the old but stable research prototype to a new commercially supported version which is planned and currently under production.

3.2.1.4 Test mapping with data

The final tests with mapping data given the previously prepared alignment mechanism could not be performed for the reasons explained above. Therefore also only a few items of the test protocol could be looked into.

Beside the problems which appeared a video documentation (<http://vimeo.com/23420503>) shows the principle usefulness of the tool. Unfortunately it could not be sufficiently tested but a later test with a stable updated version could be of interest.

3.2.2 Test results

3.2.2.1 Performance efficiency

Time behaviour: the rather large file

<https://svn.code.sf.net/p/annocultor/code/trunk/converters/vocabularies/places/EU/DE.rdf>

which includes about 120.000 triples could be downloaded and added into the triple store in less than 5 seconds when using the "load from HTTP" interface.

3.2.2.2 Compatibility

Interoperability: the vocabulary data can be added with an HTTP interface and should be in SKOS⁷ format.

3.2.2.3 Usability

The web based user interface has a clear form and navigation structure. However some improvements can be done in the style sheets of the web pages to give a clearer distinction between hyperlinks and other text.

3.2.2.4 Reliability

Recoverability: incorrect input file formats led to an error condition. This can be on one hand seen in the console output of the server which is normally not visible to the user. The web frontend responds with an erroneous page which hopefully will be corrected for the future commercial version.

3.2.2.5 Maintainability

Modularity: the installation has a clear structure with the basic software, the general framework and packages which can be added on demand. The Amalgame software is added as a package. For the future a test service for created alignments could be interesting. The adding mechanism for packages is easy to use.

Reusability: the use of a web based interface allows using the software within other system by adding a web control.

3.3 Quality assessment

3.3.1 Assessment results for VidiCert

3.3.1.1 Functional suitability

Degree to which a product or system provides functions that meet stated and implied needs when used under specified conditions

Measurement function: $FS = (X + Y) / 2$ (where X and Y are the scores computed as in the following

FS = Functional Suitability

X = Functional Completeness

Y = Functional Appropriateness

⁷ Simple Knowledge Organization System - Home Page; <http://www.w3.org/2004/02/skos/>

Interpretation of test results: FS value closer to 1 is better

Table 15 lists the Functional Completeness and Functional Appropriateness for the functionalities of VidiCert.

Functionality	Level of need	Functional Completeness	Functional Appropriateness (digitisation/migration, ingest, search/selection)
Automatic Defect Analysis Functions (categorisation based on EBU QC checks ⁸)			
Analogue Synchronisation Errors Aliases: lost lock, time-base corrector (TBC) hit, video breakup, lost video sync, horizontal distortion	Mandatory	1	1
Coloured Frames Aliases: Black Frames, Monochrome Frames, Uniform Color Frames	Mandatory	1	1
Digital Tape Dropouts Aliases: digital video tape dropout, digital hits, digital tape hits	Mandatory	0,5	1
Video Noise Aliases: image noise, noise	Desirable	1	1
Blurriness Aliases: out of focus, blur detection, sharpness	Recommended	1	1
Video Test Pattern Aliases: test card, colour bars	Mandatory	1	1
Video Field Order Aliases: field order, field dominance	Mandatory	1	1
Scanning Type Aliases: sampling, sampling structure, scanning	Mandatory	1	1
Audio Silence Aliases: mute test, minimum level	Mandatory	1	1
Audio Encoding Format Change	Recommended	1	1
General Analysis Properties			
Analysis profiles	Mandatory	1	1
No reference video required	Mandatory	1	1
Detection of multi-generation defects	Recommended	1	1
Multi-Resolution support	Mandatory	1	1
GPU support	Recommended	1	1
Interactive Validation/Verification Functions			
Check file efficiently for correct content	Mandatory	1	1
Human validation of automatic analysis functions	Mandatory	1	1
Interactive defect annotation support	Mandatory	1	1
Overall quality rating support	Mandatory	1	1
Defect severity based operation/validation	Recommended	1	1
Video output devices	Desirable	0,5	1
Individual field output	Recommended	0	0
Video output on interlaced capable devices	Desirable	0	0
Human validation during analysis phase	Desirable	0	0

Table 15: Functionality evaluation of VidiCert.

Functional Completeness (X) = 0.75

Functional Appropriateness (Y) = 1

⁸ EBU Strategic Programme on QC (EBU QC) <http://tech.ebu.ch/groups/qc>, First draft release of QC test definitions available at <http://tech.ebu.ch/docs/tech/tech3363.zip>

$$\text{Functional Suitability (FS)} = (X + Y) / 2 = 0.875$$

3.3.1.2 Performance efficiency

Performance relative to the amount of resources used under stated conditions.

Measurement function: $PE = (X+Y+Z) / 3$ where X,Y,Z are the scores computed as in the following

$$\begin{aligned} PE &= \text{Performance Efficiency} \\ X &= \text{Time behaviour} \\ Y &= \text{Resource utilization} \\ Z &= \text{Capacity} \end{aligned}$$

Interpretation of test results: PE smaller is better

$$\text{Time behaviour (X)} = 0.6 \text{ for SD Material}$$

$$\text{Resource utilization (Y)} = 0.5 + 0.03 = 0.53$$

$$\text{Capacity (Z)} = 1 / (1 + 1 + 0.5 + 1) = 0.29$$

$$\text{Performance Efficiency (PE)} = (X+Y+Z) / 3 = (0.6 + 0.53 + 0.29) / 3 = 0.473$$

3.3.1.3 Compatibility

Degree to which a product, system or component can exchange information with other products, systems or components, and/or perform its required functions, while sharing the same hardware or software environment.

Measurement function: $Co=Y$ is computed as in the following

$$\begin{aligned} Co &= \text{Compatibility} \\ Y &= \text{Interoperability} \end{aligned}$$

Interpretation of test results: Co value larger is better

$$\text{Interoperability (Y)} = (5 + 4 + 5 \text{ (all except ProRes)} + 5 + 2 + 2) / 24 = 23 / 24 = 0.958$$

$$\text{Compatibility (Co)} = Y = 0.958$$

3.3.1.4 Usability

Degree to which a product or system can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.

Measurement function: $Us = N$, where

$$\begin{aligned} Us &= \text{Usability} \\ N &= \text{Accessibility} \\ \text{For calculating the N score, see section below.} \end{aligned}$$

Accessibility (N) = 3 (Mouse, Mouse Wheel and Keyboard) / 3 = 1

Usability (Us) = N = 1

3.3.1.5 Reliability

Degree to which a system, product or component performs specified functions under specified conditions for a specified period of time.

Measurement function: $Re = L$, where

Re = Reliability

L = Recoverability

For calculating the L score, see section below.

Interpretation of test results: Re value closer to 1 is better

Recoverability (L) = $(1 + 1 + 1 + 0(\text{network interruption during single file analysis}) + 1 + 1) / 6 = 0.83$

Reliability (Re) = L = 0.83

3.3.1.6 Maintainability

Degree of effectiveness and efficiency with which a product or system can be modified by the intended maintainers.

Measurement function: $Ma = H$, where

Ma = Maintainability

H = Modularity

For calculating the H score, see section below

Interpretation of test results: Ma value closer to 1 is better

Modularity (H) = 1

Maintainability (Ma) = H = 1

Remarks: The tool supports comprehensive plugin architecture. New detectors and visualization can be plugged-in in a very modular way.

3.3.1.7 Portability

Degree of effectiveness and efficiency with which a system, product or component can be transferred from one hardware, software or other operational or usage environment to another.

Measurement function: $Po = Y$ where Y computed as in the following

Po = Portability

Y = *Installability*

Interpretation of test results: Po value closer to 1 is better

Installability (Y) = 1

Portability (Po) = Y = 1

3.3.2 Assessment results for BAVC QC Tools

3.3.2.1 Functional suitability

Degree to which a product or system provides functions that meet stated and implied needs when used under specified conditions

Measurement function: $FS=(X+Y)/2$ (where X and Y are the scores computed as in the following)

FS = Functional Suitability

X = Functional Completeness

Y = Functional Appropriateness

Interpretation of test results: FS value closer to 1 is better

Table 16 lists the Functional Completeness and Functional Appropriateness for the functionalities of BAVC QCTools.

Functionality	Level of need	Functional Completeness	Functional Appropriateness
Automatic Defect Analysis Functions (categorisation based on EBU QC checks ⁹)			
Analogue Synchronisation Errors Aliases: lost lock, time-base corrector (TBC) hit, video breakup, lost video sync, horizontal distortion	Mandatory	1	1
Coloured Frames Aliases: Black Frames, Monochrome Frames, Uniform Color Frames	Mandatory	0.5	1
Digital Tape Dropouts Aliases: digital video tape dropout, digital hits, digital tape hits	Mandatory	1	1
Video Noise Aliases: image noise, noise	Desirable	1	0,5
Blurriness Aliases: out of focus, blur detection, sharpness	Recommended	0	0
Video Test Pattern Aliases: test card, colour bars	Mandatory	1	1
Video Field Order Aliases: field order, field dominance	Mandatory	0	0
Scanning Type Aliases: sampling, sampling structure, scanning	Mandatory	0	0
Audio Silence Aliases: mute test, minimum level	Mandatory	0	0

⁹ EBU Strategic Programme on QC (EBU QC) <http://tech.ebu.ch/groups/qc>, First draft release of QC test definitions available at <http://tech.ebu.ch/docs/tech/tech3363.zip>

Functionality	Level of need	Functional Completeness	Functional Appropriateness
Audio Encoding Format Change	Recommended	0	0
General Analysis Properties			
Analysis profiles	Mandatory	0	0
No reference video required	Mandatory	1	1
Detection of multi-generation defects	Recommended	1	1
Multi-Resolution support	Mandatory	1	1
GPU support	Recommended	0	0
Interactive Validation/Verification Functions			
Check file efficiently for correct content	Mandatory	0,5 - Have to look at graphs	0,5
Human validation of automatic analysis functions	Mandatory	0	0
Interactive defect annotation support	Mandatory	0	0
Overall quality rating support	Mandatory	0	0
Defect severity based operation/validation	Recommended	0	0
Video output devices	Desirable	0,5	1
Individual field output	Recommended	1	1
Video output on interlaced capable devices	Desirable	0	0
Human validation during analysis phase	Desirable	1	1

Table 16: Functionality evaluation of QCTools.

Functional Completeness (X) = 0.29

Functional Appropriateness (Y) = 0.91

Functional Suitability (FS) = $(X + Y) / 2 = 0.6$

Remarks: The automatic defect analysis functions according to the EBU QC checks are only supported by low-level signal filters visualized as line charts. Expert knowledge is required to spot possible segments within a video by combining different filters.

In specific, these defect analysis functions require expert interpretation of a combination of filters:

- Analogue Synchronisation Errors: PSNRf, TOUT
- Coloured Frames: Detection of Black Frames only with Crop Filter
- Digital Tape Dropouts: VREP, Crop top/bottom
- Video Noise: TOUT and MSEf. Functional appropriateness is 0.5 since coarse noise (grain) cannot be detected and the charts sometimes indicate noise for clean computer generated sequences.
- Video Test Pattern: Sat, VREP, Y/U/V Max/High

3.3.2.2 Performance efficiency

Performance relative to the amount of resources used under stated conditions.

Measurement function: $PE = (X+Y+Z) / 3$ where X,Y,Z are the scores computed as in the following

PE = Performance Efficiency
 X = Time behaviour
 Y = Resource utilization
 Z = Capacity

Interpretation of test results: PE smaller is better

Time behaviour (X) = 0.7

Resource utilization (Y) = $0.125 + 0.04 = 0.165$

Capacity (Z) = 2 (C1 to C4 = 0)

Performance Efficiency (PE) = $(X+Y+Z) / 3 = (0.7 + 0.165 + 2) / 3 = 0.995$

3.3.2.3 Compatibility

Degree to which a product, system or component can exchange information with other products, systems or components, and/or perform its required functions, while sharing the same hardware or software environment.

Measurement function: $Co=Y$ is computed as in the following

Co = Compatibility
Y = *Interoperability*

Interpretation of test results: Co value larger is better

Interoperability (Y) = $(3 \text{ (MP4, MOV, AVI)} + 2 \text{ (Mpeg4 + DV)} + 3 \text{ (ProRes, Uncompressed in MOV, WMV)} + 5 + 1 + 0) / 24 = 14 / 24 = 0.583$

Compatibility (Co) = Y = 0.583

3.3.2.4 Usability

Degree to which a product or system can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.

Measurement function: $Us = N$, where

Us = Usability

N = Accessibility

For calculating the N score, see section below.

Interpretation of test results: Us value closer to 1 is better

Note: Due to resource limitations, no actual user studies can be performed.

Accessibility (N) = $2 \text{ (Mouse and Keyboard)} / 3 = 0.67$

Usability (Us) = N = 0.67

3.3.2.5 Reliability

Degree to which a system, product or component performs specified functions under specified conditions for a specified period of time.

Measurement function: $Re = L$, where

Re = Reliability

L = Recoverability

For calculating the L score, see section below.

Interpretation of test results: Re value closer to 1 is better

Recoverability (L) = $(1 + 1 + 0(\text{no warning on exit}) + 0(\text{network interrupt}) + 0(\text{termination during multi-file job})) / 5 = 0.4$ (option 6 is N/A)

Reliability (Re) = L = 0.4

3.3.2.6 Maintainability

Degree of effectiveness and efficiency with which a product or system can be modified by the intended maintainers.

Measurement function: $Ma = H$ where

Ma = Maintainability

H = Modularity

For calculating the H score, see section below

Interpretation of test results: Ma value closer to 1 is better

Modularity (H) = 1

Maintainability (Ma) = H = 1

Remarks: The tool is under constant development and functionality can thus be added.

3.3.2.7 Portability

Degree of effectiveness and efficiency with which a system, product or component can be transferred from one hardware, software or other operational or usage environment to another.

Measurement function: $Po = Y$ where Y computed as in the following

Po = Portability

Y = *Installability*

Interpretation of test results: Po value closer to 1 is better

Installability (Y) = 1

Portability (Po) = Y = 1

3.3.3 Summary of Quality Assessment Tool Evaluation

Table 17 lists the summary of the assessment for each metric for VidiCert and BAVC QCTools.

Metric	Interpretation	Assessment	
		VidiCert	QCTools
Functional Suitability	closer to 1 is better	0.86	0.6
Functional Completeness	closer to 1 is better	0.75	0.29
Functional Appropriateness	closer to 1 is better	1	0.91
Performance Efficiency	smaller is better	0.473	0.995
Time behaviour	smaller is better	0.6	0.7
Resource utilization	smaller is better	0.53	0.165
Capacity	smaller is better	0.29	2
Compatibility	closer to 1 is better	0.958	0.583
Interoperability	closer to 1 is better	0.958	0.583
Usability	closer to 1 is better	1	0.67
Accessibility	closer to 1 is better	1	0.67
Reliability	closer to 1 is better	0.83	0.4
Recoverability	closer to 1 is better	0.83	0.4
Maintainability	closer to 1 is better	1	1
Modularity	closer to 1 is better	1	1
Portability	closer to 1 is better	1	1
Installability	closer to 1 is better	1	1

Table 17: Summary of evaluation for VidiCert and BAVC QCTools.

The evaluated tools follow a different approach in regard to finding defect sections in a video: VidiCert provides specific automatic detectors for a set of archive-relevant defects. Detections need only be verified by an operator. BAVC QCTools provide a set of low-level signal filters where the detection of specific defects needs to be done by an expert user by interpreting a combination of these low-level signals.

For a concrete conclusion on the efficiency (e.g. the needed operator time) of quality assessment tools for preservation purposes, it would be necessary to do an in-depth evaluation of detection ratios, false detection ratios and user interaction time.

3.4 Technical metadata extractors

3.4.1 Test environment

For all the tests a reference machine was used with the following characteristics:
HP Workstation Z800 equipped with:

- 8 Xeon cores at 2.5 Ghz,
- 8 GB RAM,
- Linux release “Centos 6.3 64 bit”
- Windows server 2008 32 bit sevice pack 2

The technical metadata extraction tools that have been assessed during the second year of the project are:

- Mediainfo version 0.7.69
Installed as binary rpm, available for Centos 6.3

- ffprobe 2.3.2
Compiled and built directly on the machine
- MXFTechnicalMetadataExtractor 0.0.6
- MXFDump ver. 1.0.1
Compiled and built directly on the machine
- MXFAnalyzer version 2.3

Mediainfo and ffprobe work with a wide range of multimedia file formats while MXFTechnicalMetadataExtractor , MXFDump and MXFAnalyzer are specifically designed for the analysis of MXF files only. MXF (Multimedia Exchange Format) is a very important and widespread standard for the professionals like television broadcasters and video production and post-production.

All the tools were assessed with their Linux versions except for MXFAnalyzer that is commercialized only under Microsoft Windows operative systems.

3.4.2 Dataset

The audiovisual material that has been used for the tests includes all the files provided within the P4U project plus a certain amount of additional files that has been judged necessary for having a sufficient quantity and diversity of file formats. In particular were added some MPEG transport stream files (including MPEG2 and H264 essence) and some MXF files containing uncompressed essence.

The detailed composition of the dataset is shown in Table 18, differentiated by typology.

File Type	Number of file analyzed	Nbr of video streams	Nbr of audio streams	MB	Total duration
MXF-D10	8	8	8	5930	00:18:00
MXF-XDCAM	12	12	12*8 = 96	9102	00:20:02
MXF-Proxy	12	12	12*4 = 48	315	00:20:02
MXF-Uncompressed ¹⁰	11	11	12*4 = 48	15280	00:09:24
MP4	9	9	9	25168	11:02:57
MP4-H264 Proxy	20	20	20	1006	01:39:40
TS - MPEG2-SD, AVC-HD ¹¹	18	18*3 = 54 (36 MPEG2 and 18 H.264)	18*5 = 90 (72 MP2 and 18 AC-3)	94237	11:05:56
MOV – Prores	4	4	4	10918	00:11:57
OGV - flv	20	20	20	888	03:13:32
	114	150	343	162844	28:21:30

Table 18 - Composition of the data set

3.4.3 Functional suitability

3.4.3.1 Functional Completeness

Table 2 reports which metadata (e.g. video bitrate) each tool is capable to extract from multimedia files, Table 19 is similar but lists additional metadata items that apply only for MXF files and then only for MXF specialized tool: MXFTechnicalmetadataExtractor, MXFDump and MXFAnalyzer.

¹⁰ Not in the shared P4U dataset

¹¹ Not in the shared P4U dataset

Metadata Item	MedialInfo	ffprobe	MXFTechMetaExtractor	MXFDump	MXFAnalyzer
Overall bitrate mode	X				
Overall bitrate	X	X			
Video resolution	X	X	X	X	X
Frame rate mode	X				
Video frame rate	X	X	X	X	X
Video bit depth	X		X	X	X
Display Aspect ratio	X	X	X	X	X
Pixel Aspect ratio		X			
Video Scanning	X		X	X	X
Field order	X		X	X	X
Video coding	X	X	X	X	X
Video bitrate mode	X				
Video bitrate	X	X			X
GOP structure	X for MPEG2				X
Color Space	X	X			
Chroma subsampling	X	X	X	X	X
Video duration	X	X	X	X	X
Timecode	X	X		X	X
Audio nbr of channels	X	X	X	X	X
Audio sample rate	X	X	X	X	X
Audio bit depth	X	X	X	X	X
Audio coding	X	X		X	X
Audio bitrate mode	X				
Audio bitrate	X	X			X
Audio duration	X	X	X	X	X

Table 19 – Generic extraction functions provided by the tools

Metadata Item	MXFTechMetaExtractor	MXFDump	MXFAnalyzer
Operational pattern	X	X	X
ActiveFormatDescriptor	X	X	X
Header partition status	X	X	X
Footer partition status		X	X
Essence container mapping	X	X	X
Stored width and height	X	X	X
Sampled width and height	X	X	X
Display width and height	X	X	X

Table 20 - MXF specific extraction functions provided by the tools

According to the assessment template and to Table 2 and Table 3, the functional completeness of the tools is easily calculated as the ratio of extracted metadata items and the overall number of them:

MedialInfo = 23/25 = **0.92**

ffprobe= 17/25 = **0.68**

MXFTechnicalMetadataExtractor=20/33 = **0.606**

MXFDump= 23/33 = **0.697**

MXFAnalyzer=26/33 = **0.788**

3.4.3.2 Functional Correctness

Assumed that a certain tool treats a specific metadata item, it is not certain that for every file format and for every single file instance, the extraction is made correctly. The role of functional correctness is exactly to assess this kind of errors.

Table 21, Table 23, Table 25, Table 27 show for each tool, the number of errors encountered sub-divided by file format and with a short description of the error. Table

22, Table 24, Table 26, Table 28 report the correctness indexes for each metadata whose extraction is affected by some error (correctness lower than 1).

File Type	Error	Error rate	Description
MXF-D10	Video bitrate	6/8	In 6 cases the video bitrate had not been correctly detected and declared as N/A.
	Overall bitrate	6/8	When video bitrate is not detected correctly the overall bitrate is made to coincide with the audio bitrate
	Pixel aspect ratio	8/8	Pixel aspect ratio takes into consideration all the 608 lines while only 576 should be considered
	Color space	8/8	Not detected (set to unknown)
MXF-XDCAM	All ok		
MXF-Proxy	Video bitrate	12/12	Not detected at all (set to N/A)
	Color space	12/12	Not detected (set to unknown)
MXF-Uncompressed	Video bitrate	11/11	Not detected at all (set to N/A)
	Color space	11/11	Not detected (set to unknown)
	Timecode	11/11	Not detected at all (set to N/A)
MP4	Color space	9/9	Not detected (set to unknown)
	Audio bit depth	9/9	Wrongly detected to 0
MP4-H264 Proxy	Color space	20/20	Not detected (set to unknown)
	Audio bit depth	20/20	Wrongly detected to 0
TS - MPEG2-SD, AVC-HD Note: files from the same transport stream where there are 2 MPEG2 SD with 2 separated audio stereo channels and one H264 with a single stereo AC-3 audio.	Video Duration	(10+7)/(18*3)	Wrongly detected (N/A) in 10 cases for MPEG2 and in 7 cases for H.264
	Audio Duration	(34+8)/(18*5)	Wrongly detected (N/A) in 34 cases for MPEG Layer 2 audio (MP2) and in 8 cases for ATSC A/52A (AC-3).
	Overall bitrate	14/18	Wrongly detected in 13 cases (declared a lower bitrate).
	Video bitrate	18/18	Not detected (N/A)
	Pixel Aspect Ratio	18/(18*3)	In each file PAR is wrong in one video channel over 3. It happens with a channel with resolution 704x576, real par should be 176:81
	Color Space	(18*2)/(18*3)	In each file the color space is wrong for 2 of the three video channels. It is correct only for the H264 and wrong for MPEG2 programmes.
	Timecode	18/(18*3)	Not detected (N/A) for the H.264 channels
	Audio nbr channels	18/(18*5)	In each file the color space is wrong for the AC-3 channel (associated with the H264) for which it is signaled 6 channels while there are only 2.
	Audio bit depth	(18*5)/(18*5)	Always wrongly signaled as 0
MOV	Display aspect ratio	1 / 4	Declared as 0:1, non a valid value
	Pixel aspect ratio	1 / 4	Declared as 0:1, non a valid value
	Color space	3 / 4	Not detected (unknown)
OGV – flv	Video bit rate	20 / 20	Non detected (N/A)
	Display aspect ratio	20 / 20	Declared as 0:1, not a valid value

	Pixel aspect ratio	20 / 20	Declared as 0:1, not a valid value
	Audio bit depth	20 / 20	Wrongly detected to 0

Table 21 - FFprobe, detail of the errors

Extracted metadata	Correctness
Overall bitrate	1- 20 / 114 = 0.825
Video bitrate	1- 68 / 150 = 0.547
Display aspect ratio	1- 22 / 150 = 0.853
Pixel Aspect ratio	1- 48 / 150 = 0.680
Video Duration	1- 17 / 150 = 0.887
Color space	1- 99 / 150 = 0.340
Audio Duration	1- 42 / 343 = 0.878
Timecode	1- 29 / 10112 = 0.713
Audio bit depth	1- 120 / 343 = 0.650
Audio nbr of channels	1- 18 / 343 = 0.950

Table 22 - FFprobe, calculation of correctness indexes

File Type	Error	Error rate	Description
MXF-D10	Overall bitrate	6/8	Not declared in 6 cases, it happens when no duration is written in metadata
	Overall bitrate mode	8/8	Not declared
	Video bitrate mode	8/8	Not declared
	Video Duration	6/8	Non declared
	Audio Duration	6/8	Non declared
MXF-XDCAM	All ok		
MXF-Proxy	Overall bitrate mode	12/12	Not declared
	Video bitrate mode	12/12	Not declared
	Video bitrate	12/12	Not declared
	Chroma subsampling	12/12	Not declared
	Audio bitrate mode	12/12	Not declared
	Audio bitrate	12/12	Not declared
MXF-Uncompressed	Video bitrate mode	11/11	Not declared
MP4	Color space	9/9	Not declared
	Video bitrate mode	9/9	Not declared
	Audio bit depth	9/9	Not declared
MP4-H264 Proxy	Video bitrate mode	20/20	Not declared
	Audio bit depth	20/20	Not declared
TS - MPEG2-SD, AVC-HD Note: files from the same transport stream where there are 2 MPEG2 SD with 2 separated audio stereo channels and one H264 with a single stereo AC-3 audio.	Video bitrate	(18*3)/(18*3)	Both MPEG2 and H264 are declared as variable bit rate but the average bitrate is not declared (only the maximum)
	Frame rate mode	(18*3)/(18*3)	Not declared
	Timecode	18/(18*3)	Not declared for AVC (yes for MPEG2)
	Audio nbr of channels	15/(18*5)	Wrongly declared 6 channels for AAC in 15 cases

12 A subset of the data set files does not support or include the timecode (e.g. MP4 and OGV)

	Audio bit depth	(18*4)/(18*5)	Not declared for MPEG2 (correctly declared for AAC)
MOV	Overall bit rate mde	3/4	Declare correctly in the sample with video prores, otherwise not declared
	Video bitrate mode	1 / 4	Not declared for the sample with AVC video inside, otherwise correctly detected
	Video bit depth	2 / 4	Not declared for the sample with prores and the sample with uncompressed NTSC
	Video Scanning	2 / 4	Not declared in the sample with uncompressed essence
	Field Order	4 / 4	Not declared
OGV - flv	Video bitrate mode	20/20	Not declared
	Frame rate mode	20/20	Not declared
	Video Scanning	20/20	Not declared
	Field order	20/20	Not declared
	Color space	20/20	Not declared
	Chroma subsampling	20/20	Not declared
	Video bit depth	20/20	Not declared
	Audio bit depth	20/20	Not declared

Table 23 – MedialInfo, detail of the errors

Extracted metadata	Correctness
Overall bitrate mode	1-23/114 = 0.798
Overall bitrate	1-6/114 = 0.947
Video bitrate mode	1-81/150 = 0.460
Video bitrate	1-66/150 = 0.560
Video duration	1-6/150 = 0.960
Chroma subsampling	1-32/150 = 0.787
Color space	1-29/150 = 0.807
Frame rate mode	1-74/150 = 0.507
Timecode	1-18/101 = 0.822
Video bit depth	1-22/150 = 0.853
Video scanning	1-22/150 = 0.853
Field order	1-24/150 = 0.840
Audio bitrate mode	1-12/150 = 0.920
Audio bitrate	1-12/150 = 0.920
Audio nbr of channels	1-15/343 = 0.956
Audio duration	1-6/343 = 0.983
Audio bit depth	1-121/343 = 0.647

Table 24 – Mediainfo, calculation of correctness indexes (where some error occurred)

File Type	Error	Error rate	Description
MXF-D10	Audio coding	8/8	Not detected ("null")
	Duration	6/8	Not detected ("-1")
	Active format description	8/8	Not detected ("null")
MXF-XDCAM	Audio coding	12/12	Not detected ("null")
	Active format description	12/12	Not detected ("null")
MXF-Uncompressed	Audio coding	11/11	Not detected ("null")
	Active format description	11/11	Not detected ("null")
MXF-Proxy	Exception	12/12	

Table 25 – MXFTechnicalMetadataExtractor, detail of errors

Extracted metadata	Correctness
Duration	1 - 18/43 = 0.581
Audio coding	1 - 43/43 = 0
Active format description	1 - 43/43 = 0
All the others	1 - 12/43 = 0.721

Table 26 - MXFTechMetaExtractor, calculation of correctness indexes

File Type	Error	Error rate	Description
MXF-D10	Active format description	8/8	Not declared. Despite these files do not have AFD, it is considered an error because not declared.
MXF-XDCAM	Active format description	12/12	Not declared. Despite these files do not have AFD, it is considered an error because not declared.
MXF-Uncompressed	Active format description	11/11	Not declared. Despite these files do not have AFD, it is considered an error because not declared.
MXF-Proxy	Active format description	12/12	Not declared. Despite these files do not have AFD, it is considered an error because not declared.

Table 27 – MXFDump, details of errors

Extracted metadata	Correctness
Active format description	1-43/43 = 0

Table 28 – MXFDump, calculation of correctness indexes

Table 29 summarizes the correctness of each tool with respect of each metadata item considered. The last row reports the correctness score per tool as an average over all the metadata item that the specific tool is able to extract.

Metadata Item	MedialInfo	ffprobe	MXFTechMetaExtractor	MXFDump	MXFAnalyzer
Overall bitrate mode	0.798				
Overall bitrate	0.947	0.825			
Video resolution	1	1	0.625	1	1
Frame rate mode	0.507				
Video frame rate	1	1	0.625	1	1
Video bit depth	0.853		0.625	1	1
Display Aspect ratio	1	0.853	0.625	1	1
Pixel Aspect ratio		0.680			
Video Scanning	0.853		0.625	1	1
Field order	0.840		0.721	1	1
Video coding	1	1	0.721	1	1
Video bitrate mode	0.460				
Video bitrate	0.560	0.547			1
GOP structure	1				1
Color Space	0.807	0.340			
Chroma subsampling	0.787	1	0.721	1	1
Video duration	0.960	0.887	0.581	1	1
Timecode	0.822	0.713		1	1
Audio nbr of channels	0.956	0.950	0.721	1	1
Audio sample rate	1	1	0.721	1	1
Audio bit depth	0.647	0.650	0.721	1	1
Audio coding	1	1	0	1	1
Audio bitrate mode	0.920				
Audio bitrate	0.920	1			1
Audio duration	0.983	0.878	0.581	1	1
Operational pattern			0.721	1	1
ActiveFormatDescriptor			0	0	1
Header partition status			0.721	1	1
Footer partition status			0.721	1	1
Essence container mapping			0.721	1	1
Stored width and height			0.721	1	1
Sampled width and height			0.721	1	1
Display width and height			0.721	1	1
GLOBAL CORRECTNESS	0.859	0.843	0.621	0.957	1

Table 29 - Tools correctness calculated over the given dataset

3.4.4 Performance efficiency

3.4.4.1 Time Behaviour

All the metadata extractions have been made in sequence on the reference system while measuring the elapsed time and the use of the resources in terms of CPU and Memory. For doing that we used the command “*time*” easily available under the used Linux distribution.

We found that all the inspections, with the only exception of those made with MXFAnalyzer, took less than half a second even for larger files (more than 9 GB). A simple read of such a file took around 1 minute and 20 seconds on the used system and this proves that the considered tools do not analyze the entire file but rather they gather information from either the header, the footer or other areas where technical metadata is declared within the media file.

MXFAnalyzer on the other hand takes more time because it goes down to the bit-stream and analyze all the MXF partitions and the KLV structure of the those files. Execution time is then not directly comparable with that of the other examined tools.

Figure 1 shows the execution time (in hundredths of seconds) in relation with the analyzed files and their size in megabytes. It can be easily seen that tools are all very fast and that in general there is not direct relation between size and elapsed time, rather each tool has its own execution time which is roughly constant.

With MXFAnalyzer instead there is a linear relation between the file size and the elapsed time as Figure 2 clearly shows. For bigger files it took around 50 seconds to complete that is less than pure read of the entire file (1 minutes and 20 seconds).

For this part of the assessment it is considered not reasonable to use the formula proposed in the assessment template due to negligible execution time.

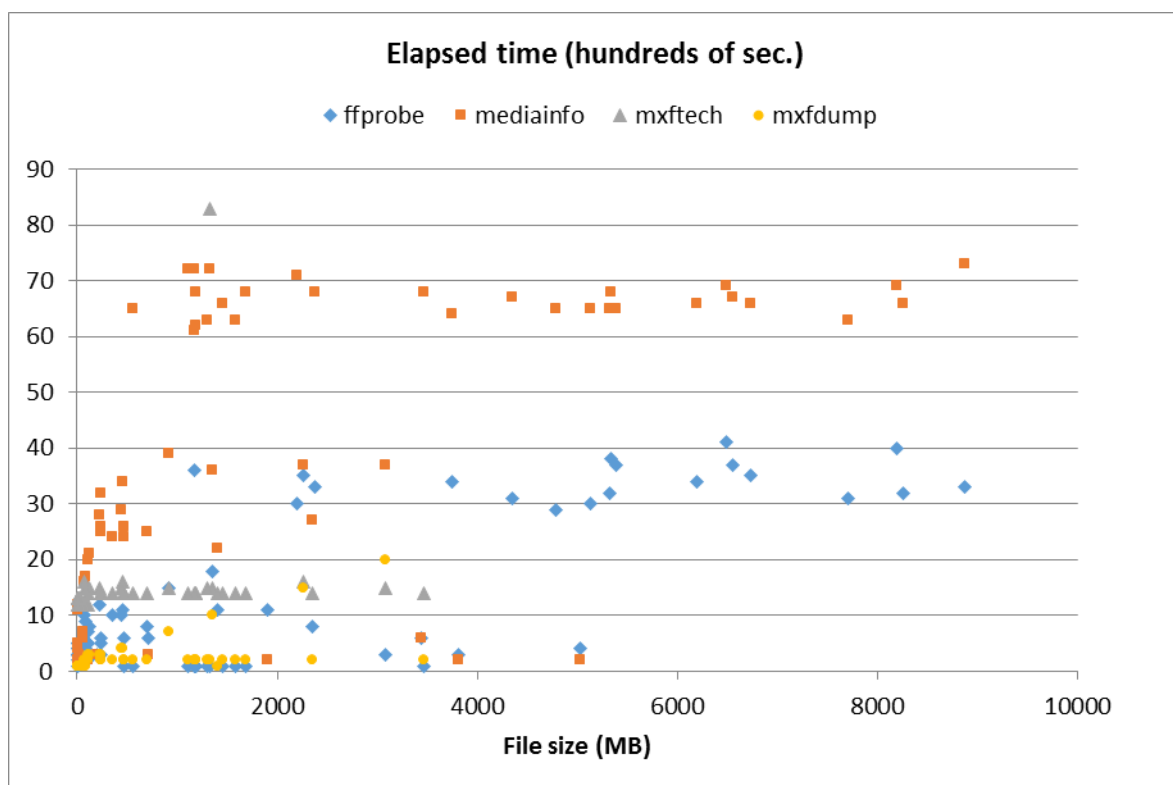


Figure 6 – Analysis time by file and tool

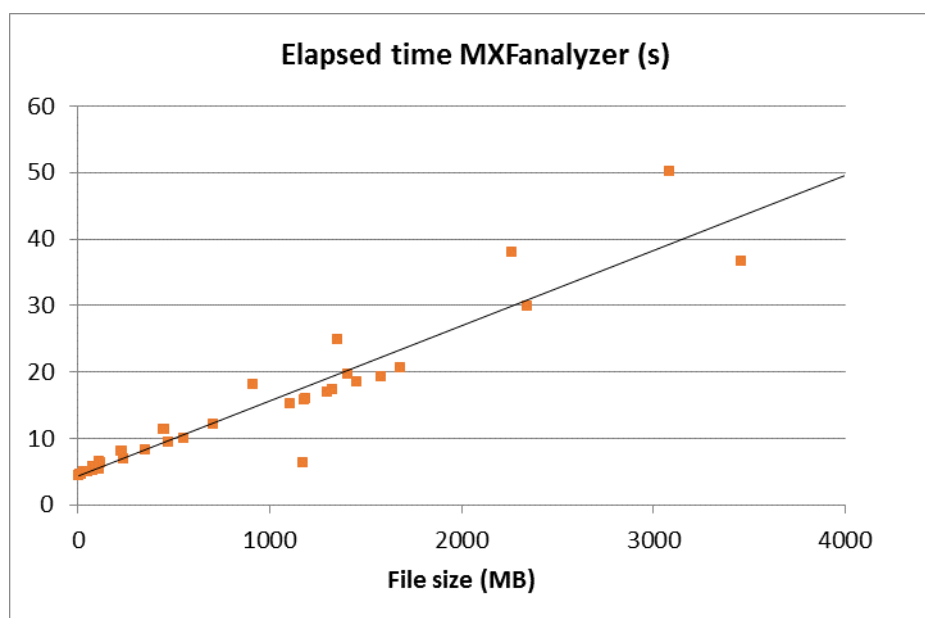


Figure 7 - Analysis time by file for MXFAnalyzer

For completeness Table 30 reports the average elapsed times measured during the execution of the tests.

	Mean elapsed time (sec.)
FFprobe	0.098
Mediainfo	0.257
MXFTechMetaExtractor	0.155
MXFDump	0.029
MXFAnalyzer	12.70

Table 30 - Mean elapsed time for each tool

3.4.5 Resource Utilization

Resource utilization is limited in time because of very quick analysis, moreover also the percentage amount of used resources is quite low.

Figure 3 shows for all the tools, except MXFAnalyzer, the CPU usage (fraction of CPU cores) in relation with the analyzed files and their size in Mbytes. When an entire core is used the graph assumes the value 1, it is then clear that FFprobe, Mediainfo and MXFDump use only one core (our machine dispose of 8) while MXFTechMetadataExtractor (written in Java) automatically scales on more cores when available even though it was not written specifically with a multithread approach.

In general it is not noticeable a relation between file size and resource utilization, rather the usage is roughly constant.

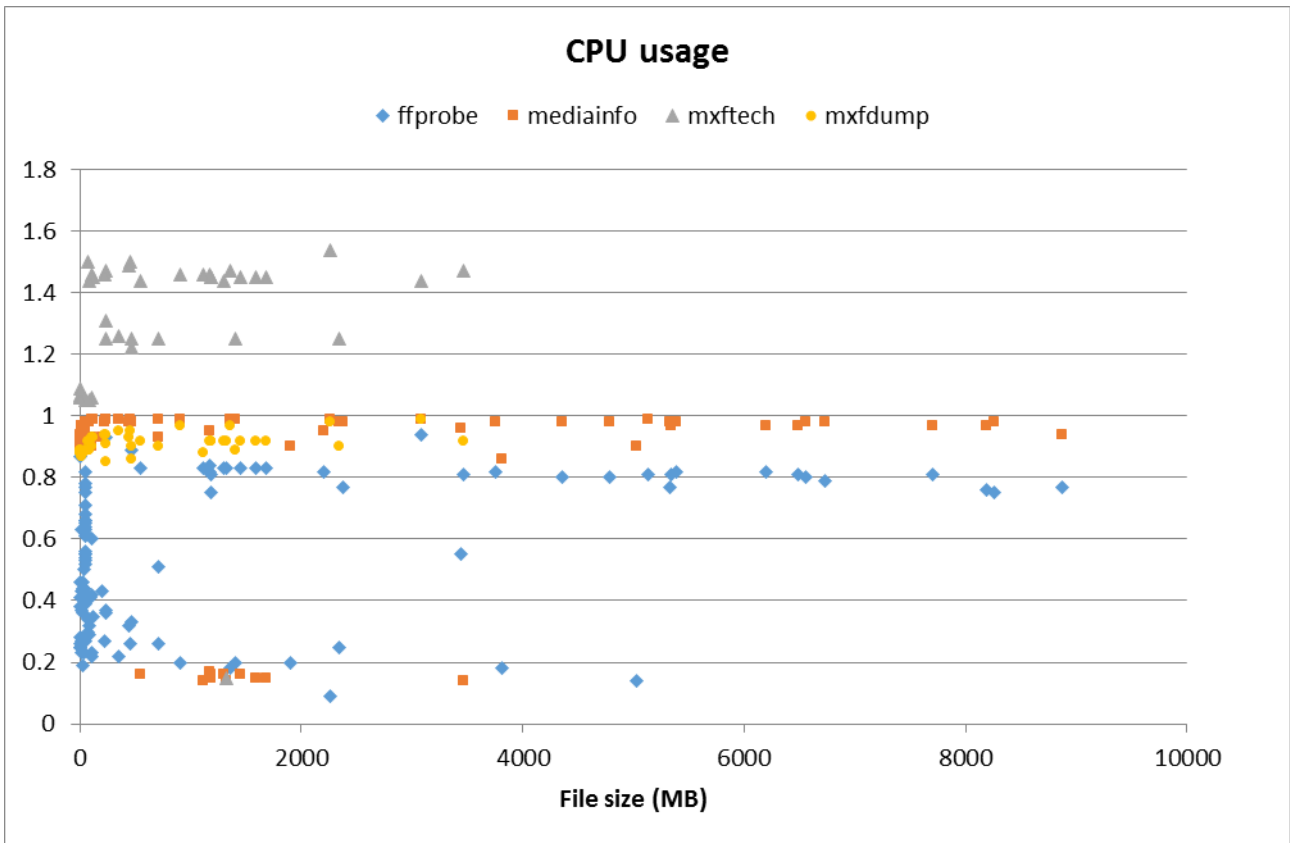


Figure 8 – CPU usage by file and tool

Figure 4 shows the CPU measure for MXFAnalyzer, all the 43 MXF files has been analyzed in around 11 minutes and the overall CPU usage is on average around 11%. In relation to the graph of the other tools (Figure 9) this means a value of around 0.9 (considering the 8 cores of the reference system). The software is multithreaded and spans on more cores as is visible in a snapshot taken during the elaborations in Figure 10.

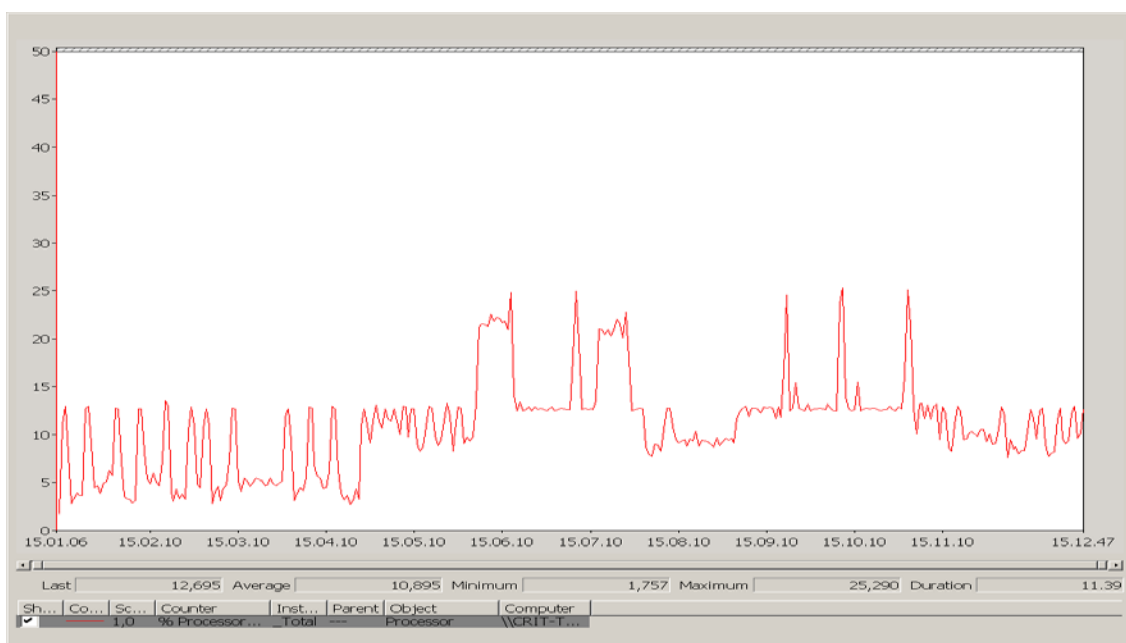


Figure 9 - MXFAnalyzer, overall CPU usage

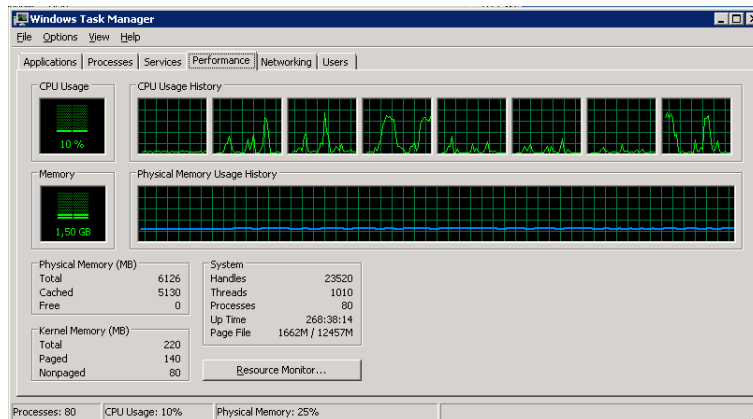


Figure 10 - MXFAnalyzer CPU usage per core

Figure 11 shows the average use of the RAM memory for each tool and for each analyzed file. RAM occupation is low and always under 200 MB for all the considered tools.

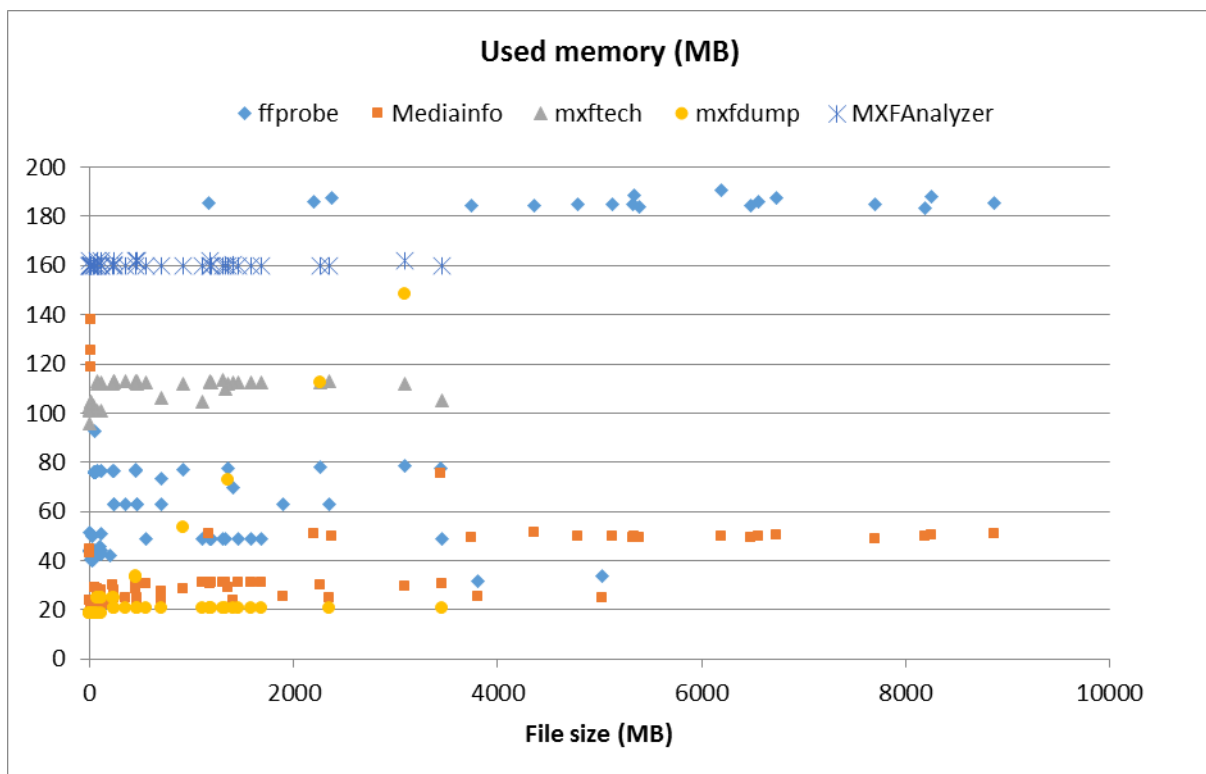


Figure 11 – Used memory by file and tool

Table 31 reports the average CPU and Memory usage measured during the execution of the tests and the global Resource Utilization according to the formula provided in the assessment template.

	Avg. CPU	Avg. Memory (MB)	Resource utilization (Avg.CPU/8 + Avg.Memory/8000) / 2
FFprobe	54 %	81	3.9%
Mediainfo	87 %	34	5.7%
MXFTechMetaExtractor	128 %	109	8.7%
MXFDump	91 %	28	5.9%
MXFAnalyzer	88%	160	6.5%

Table 31 – Mean resource usage for each tool

3.4.6 Compatibility (Interoperability)

According to the proposed assessment template we consider two aspects for calculating the Interoperability.

The first one is related to the terminology used for the metadata, more or less all the considered tools adopt the current in use terminology or a well known alias. Few exceptions are for ffprobe that uses the not very common term “sample aspect ratio” on behalf of “pixel aspect ratio” and the term “px_fmt” i.e. *pixel format* for expressing both the chroma subsampling and the color space. For this reason the score assigned to ffprobe is 0.94 rather than 1.

The second aspect considers system integration, Table 32 reports for each tool which integration method is actually supported.

	System call	Web service or REST	SDK (library)	B
FFprobe	yes	no	yes	0.66
Mediainfo	yes	no	yes	0.66
MXFTechExtractor	yes	no	yes	0.66
MXFDump	yes	no	yes	0.66
MXFAnalyzer	yes	yes	yes	1

Table 32 – Possible ways for tool integration

Table 33 reports the Compatibility for each tool as an average of the two considered aspects.

	A	B	Y=(A + B)/2
FFprobe	0.94	0.66	0.8
Mediainfo	1	0.66	0.83
MXFTechExtractor	1	0.66	0.83
MXFDump	1	0.66	0.83
MXFAnalyzer	1	1	1

Table 33 - Compatibility for each tool

3.4.7 Usability

Usability is expressed with three sub-characteristics: Operability, User error protection and User Interface aesthetics which are detailed in the following sub-chapters 1.6.1, 1.6.2 1 and 1.6.3. Table 34 reports each values and the overall usability in the rightmost column.

	K Operability	L User error prot.	M GUI aesthetics	Us=(K + L + M)/3
FFprobe	0.5	0.25	0	0.25
Mediainfo	1	0.5	0.66	0.72
MXFTechExtractor	0.5	0.25	0	0.25
MXFdump	0.5	0.25	0	0.25
MXFAnalyzer	1	0.75	0.42	0.72

Table 34 - Usability of tools

3.4.7.1 Operability

According to the assessment template, two aspects are considered. The first regards the availability of an integrated help (column A), the other the availability of an XML export (column B). Table 35 reports what was observed and the resulting Usability score.

	A	B	K=(A + B)/2
FFprobe	1	0	0.5
Mediainfo	1	1	1
MXFTechExtractor	1	0	0.5
MXFdump	1	0	0.5
MXFAnalyzer	1	1	1

Table 35 – Operability for each tool

3.4.7.2 User error protection

According to the assessment template, two aspects are considered. The first tells whether the GUI uses - where applicable - controlled vocabularies for the fields (column A), the other whether there is a formal check of the input parameter values or in alternative a precise reporting of the problem with inputs (column B).

Table 36 reports what was observed and the resulting User error protection score.

	A	B	K=(A + B)/2
FFprobe	0	0.5	0.25
Mediainfo	0.5	0.5	0.5
MXFTechExtractor	0	0.5	0.25
MXFdump	0	0.5	0.25
MXFAnalyzer	1	0.5	0.75

Table 36 – User error protection for each tool

3.4.7.3 User interface aesthetics

In this chapter is analyzed the effectiveness and appealing of the GUI (graphical user interface) when available. Only Mediainfo and MXFAnalyzer come with a native GUI, others like ffprobe can be complemented with third party graphical front-end but it is not assessed here. According to the assessment template several aspects are considered, the meaning of which is recalled here with reference to columns in Table 20.

A: Language configurability

B: Color configurability

C: Customization of the disposition of fields

D: Input section completeness

E: Output display configuration

F: Presence of an integrated 'help'

Table 37 reports what was observed and the resulting scores, zero means that the functionality is not available at all, 0.5 that the feature is available but with limited functionality.

	A	B	C	D	E	F	$K=(A+B+C+D+E+F)/6$
Mediainfo	1	0	1	0.5	1	0.5	0.66
MXFAnalyzer	0	0	0.5	0.5	0.5	1	0.42

Table 37 – User interface aesthetics for each tools having a GUI

3.4.7.4 Reliability (Maturity)

The Reliability of the tools is provided as an estimation of the technology readiness level (TRL) as defined in D3.1, chapter 1.3. The highest score is for MXFAnalyzer because it claims use in production scenarios either directly or as a part of other software.

Mediainfo has also quite a good TRL because it is used by professional software solutions from Digimetrics¹³.

	TRL
FFprobe	6
Mediainfo	7
MXFTechExtractor	5
MXFdump	5
MXFAnalyzer	9

Table 38 – Maturity of each tool

3.4.7.5 Maintainability (Modifiability)

According to the assessment template, two aspect are considered. The first considers if the software is open source or not (column A), the other if the software is currently maintained and in evolution (column B).

	A	B	$K=(A+B)/2$
FFprobe	1	1	1
Mediainfo	1	1	1
MXFTechExtractor	1	0.5	0.75
MXFdump	1	0.5	0.75
MXFAnalyzer	0	1	0.5

Table 39 – Maintainability of the tools

3.4.7.6 Portability (Installability)

According to the assessment template, two aspects are considered. The first relates to the availability of an installer or an installation procedure (column A), the other to the fact that the installation effectively explains the encountered problems (column B).

	A	B	$K=(A+B)/2$
FFprobe	1	0.5	0.75
Mediainfo	1	0.5	0.75
MXFTechExtractor	1	0.5	0.5

¹³ <http://digi-metrics.com/> they provide software for file-based test and measurement

MXFdump	0.5	0.5	0.5
MXFAnalyzer	1	1	1

Table 40 – Portability of the tools

3.5 Preservation Platforms Evaluation

In this section we report the results of the assessment for three digital preservation platforms (DSpace, Archivematica, RODA), performed according to the measurement plan described above. For each of the three platforms, we both provide a score for each assessment criteria and sum up all the results into a summary table. The test environment and the dataset are described here below.

The assessment of the platform has been performed within a test environment at EURIX. We used a dedicated server with a virtualization environment e for each platform we have setup a separate virtual machine with the minimal hardware and software requirements available in the documentation.

The test infrastructure includes a dedicated DELL PowerEdge R320 server, equipped with Linux (Ubuntu 12.04 LTS 64-bit), and an 8 TB NAS for data storage, connected via dedicated Gb Ethernet connection. The testbed server provides a virtualization environment based on Kernel-based Virtual Machine (KVM). KVM is a full virtualization solution for Linux on x86 hardware containing virtualization extensions (Intel VT or AMD-V) and consists of a loadable kernel module providing the core virtualization infrastructure and a processor specific module.

Each platform to be tested has been deployed in a separate virtual machine, with a private virtualized hardware: a network card, disk, graphics adapter, etc. The kernel component of KVM is included in the main Linux kernel starting from version 2.6.20. Virtual disks can be converted to/from other virtualization formats to be used with other virtualization solutions, such as VirtualBox, VMWare or XEN. This feature is quite helpful since for some preservation platforms a virtual machine with a pre-configured installation of the platform is available from the official project page. We followed the available documentation for each platform and we built each one from scratch, but we used this pre-configured VMs for preliminary tests and for comparison with the new built platform, mainly to check configuration.

The hardware and software configuration for each VM is reported in the table below:

Platform	Hardware Requirements		Prerequisite Software
Dspace (v4.1)	minimal	2 GB RAM, 20 GB HDD	<ul style="list-style-type: none"> • Unix-like OS or Microsoft Windows • Oracle Java JDK 7 or OpenJDK 7 • Apache Maven 3.x • Apache Ant 1.8 • Relational Database • Servlet 3.0 Container (Tomcat 7+ or Jetty 8+)
	mid-range	4 GB RAM, 200 GB HDD	
	high-end	8 GB RAM, Quad Core, 73 GB 15,000 rpm network disks in RAID gigabit	
Archivematica (v1.2)	small-scale	2 GB RAM, 7 GB HDD, Dual	1. Ubuntu 12.04 LTS*

		Core	
	production	8 GB RAM, 10 GB HDD, Dual Core i5	
RODA (v1.1.0)**	not specified in the documentation***		<ul style="list-style-type: none"> • Unix-like OS or Microsoft Windows • Oracle Java JDK 7 or OpenJDK 7 • Relational Database • Servlet 3.0 Container (Tomcat 7+ or Jetty 8+) • Maven 2

Table 41: Hardware and Software Configuration

*Since version 1.0, *Archivematica* installation makes use of *Ubuntu apt-get* tool to retrieve *Archivematica* packages and other dependencies from *Ubuntu* repositories.

**RODA is built on top of *Fedora Commons*. The reported software requirements are referring to *Fedora* (v3.8.0).

***For this platform the same configuration as for *DSpace* has been adopted.

The three platforms under assessment virtually support any kind digital content: images, text documents, audio files and videos.

Concerning the test dataset, we used the AV files provided by project partners in the Presto4U dataset. For preliminary test during configuration we also used other content types such as images and text documents, for quick tests.

In the following sections we provide the details of the assessment for each of the three platforms.

3.5.1 DSpace

3.5.1.1 Functional Suitability

Degree to which a product or system provides functions that meet stated and implied needs when used under specified conditions.

Measurement function: $FS = (X+Z)/2$ where

FS = Functional Suitability

X = Functional Completeness

Z = Functional Appropriateness

X indicates how complete the implementation according to requirement specifications is.

As reported in Section 2.1.1 of D3.2, X can be calculated as follows:

$$X = (X1+X2*0.5+X3*0.25)/1.75 = (0.889+0.900*0.5+0.750*0.25)/1.75 = 0.872.$$

Where $X1=1-(1/9)$, $X2= 1-(1/10)$ and $X3= 1-(1/4)$. A is the number of missing or unsatisfying mandatory functions, B is the number of mandatory functions assessed in the evaluation, C is the number of missing or unsatisfying recommended functions assessed in the evaluation, D is the number of recommended functions, E is the number of missing or unsatisfying desirable functions assessed in the evaluation and F is the number of desirable functions. In particular, with respect to the table reported in Section 2, among the mandatory functions the content quality control (M9) was missing in the default installation. The functions R4 among the recommended and the D3 among the desirable ones were missing, too. Those features may be available as plug-ins or external tools.

Z describes how many functions with no problems are implemented for the appropriate functions for pursuing a specific task. Z can be calculated as:

$$Z = A/B = 23/26 = 0.885.$$

Where A is the sum of the scores of the implemented functions and B is the total amount of implemented functions.

$$\text{Therefore } FS = (X+Z)/2 = 0.879.$$

Interpretation of test results: FS value closer to 1 is better. The list of functions for the preservation platform considered during the assessment is presented in Section 2.5.1 of D3.2.

3.5.1.2 Performance Efficiency

Measurement function: $PE = Z$ where

PE = Performance Efficiency

Z = Capacity

Useful element for the evaluation of the capacity can be: the number of requests or simultaneous access per unit of time; the number of simultaneous jobs accepted in the ingestion queue or the number of tasks executed in parallel during a preservation workflow.

Such elements are strictly related to the hardware of the system into which the platforms are executed. For instance, since it is common for a new job or online request to generate a new thread, the availability of several computational units would improve the operation time of the platforms.

Due to the previous considerations, if the platform architecture allows a uniform distribution of the tasks, the capacity is scalable and thus the platform should get a good evaluation.

Since DSpace has no explicit limitations concerning the number of tasks that can be managed in parallel, the hardware and the underlying technology determine the operational time of the system. Moreover DSpace is adopted in production environments where a significant number of concurrent tasks may be required. Several publications (books and articles) provide information about testing the scalability of a DSpace-based archive, DSpace has been tested with millions of items representing different content types. The DuraSpace community supporting and maintaining DSpace continuously improves DSpace software to fix memory leaks and other issues affecting DSpace performances. Finally, the building blocks of DSpace are maintained by wide communities (e.g. Tomcat) and such components can be tuned in the most suitable way. For these reasons it seems reasonable to assign the following score:

$$Z = 1.000.$$

Therefore PE = 1.000.

Interpretation of test results: PE closer to 1 is better.

3.5.1.3 Compatibility

Measurement function: $Co = (X+Y)/2$ where

$Co =$ **Compatibility**

$X =$ Co-existence

$Y =$ Interoperability

As explained in Section 2.1.3 of D3.2, X indicates how flexible is the product in sharing its environment with other products without adverse impacts on other products.

It is possible to evaluate if the platform requires an exclusive usage of a component such as the database. In case the database can be shared among other systems, the platform should get a good score for this feature (between 0 and 1).

DSpace lets the user choose between several databases and to access an existing one, thus the following score should be assigned:

$$X = 1.000.$$

Y indicates how accurately is implementation of data exchange format determined between linking systems. It can be expressed as:

$$Y = A/B = 0.833.$$

Where $A = 5$ is the number of formats into which data can be exported in order to be exchanged with other platforms. $B = 6$ is the total number of data exportation formats provided by the platforms being assessed. The formats taken into account for this evaluation include among others PREMIS, DublinCore, METS and Simple Archive Format, just to name a few. For the list of supported format please refer to the documentation.

$$\text{Therefore } Co = (X+Y)/2 = 0.917.$$

Interpretation of test results: Co value closer to 1 is better.

3.5.1.4 Usability

Degree to which a product or system can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.

Measurement function: $Us = (K+L)/2$ where

$Us =$ **Usability**

$K =$ Operability

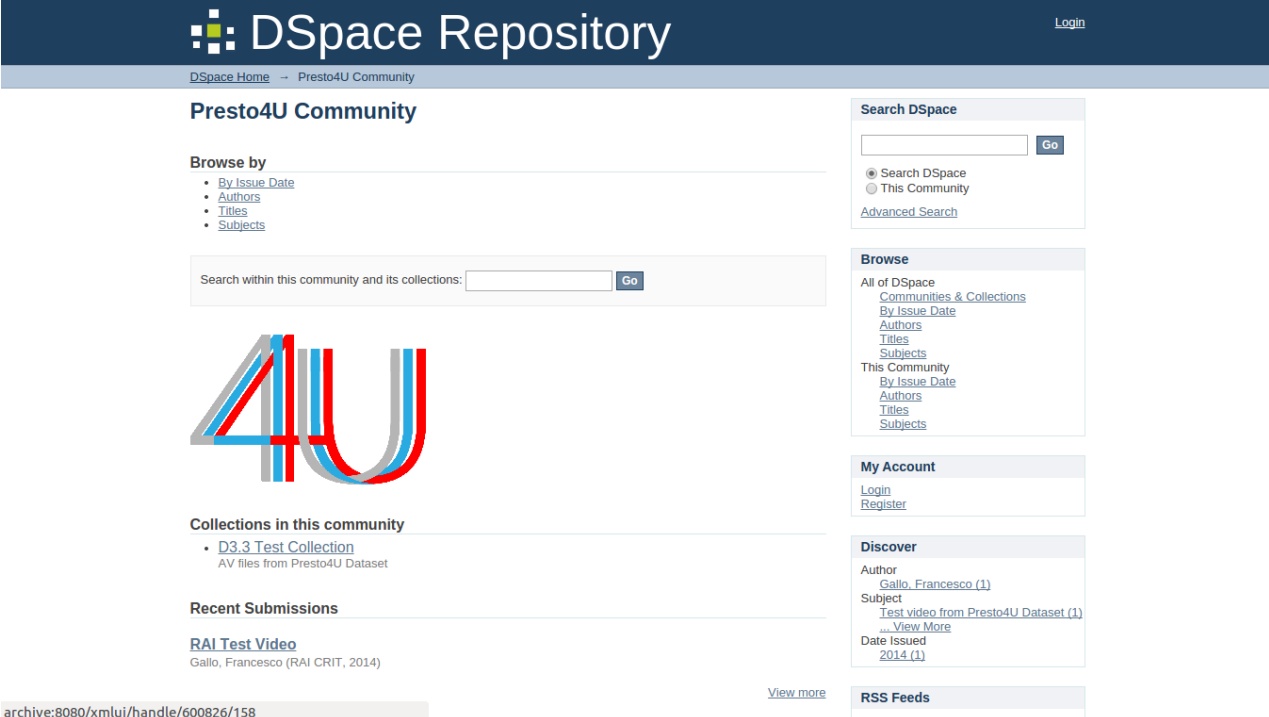
$L =$ User error protection

According to its definition, the operability indicates the degree to which the platform has attributes that make it easy to operate and control. A good estimation of K may come from the evaluation of the user interface provided by the platform. In case a clear and intuitive interface is provided the platform should get a good mark (between 0 and 1).

We tested the GUI provided by DSpace when performing the most common preservation functions such as the ingest, the access and the archive administration. We tried typical producer and consumer tasks with the AV files available in the dataset and also tried administrative tasks such as archive monitoring. The DSpace GUI comes in two flavours, a JSP-based interface and a lightweight XML-based interface enabling the usage on different client desktops and with the most popular browsers. After this test, we assign a relative score between the platforms under assessment. Since the GUI provided by DSpace was clear and intuitive, a reasonable score could be:

$$K = 0.800.$$

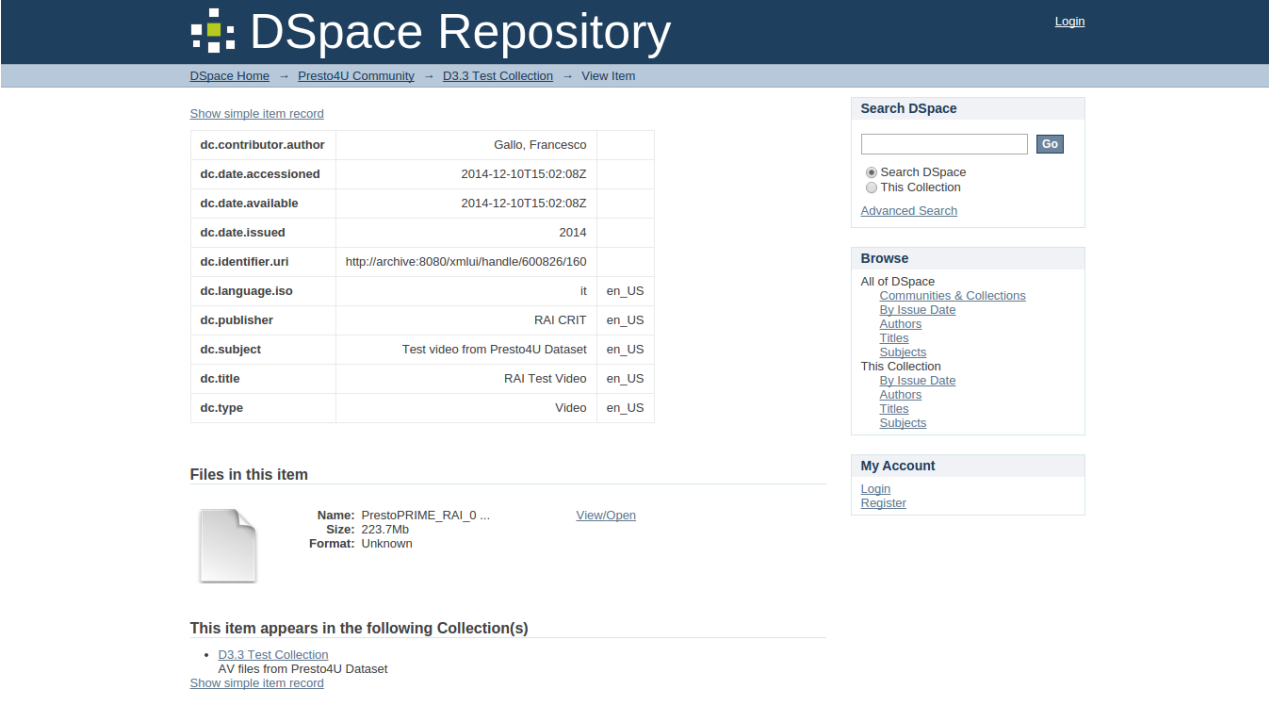
Figure 12 and Figure 13 below, show an example of the DSpace GUI through which a test collection has been created.



The screenshot shows the DSpace Repository interface for the Presto4U Community. The header includes the DSpace logo and a 'Login' link. The breadcrumb trail is 'DSpace Home -> Presto4U Community'. The main content area is titled 'Presto4U Community' and features a 'Browse by' section with links for 'By Issue Date', 'Authors', 'Titles', and 'Subjects'. Below this is a search box for the community and its collections. A large '4U' logo is displayed. The 'Collections in this community' section lists 'D3.3 Test Collection' with a sub-description 'AV files from Presto4U Dataset'. The 'Recent Submissions' section shows 'RAI Test Video' by Gallo, Francesco (RAI CRIT, 2014). A 'View more' link is present. The right sidebar contains a 'Search DSpace' box, a 'Browse' section with links for 'All of DSpace' and 'This Community', a 'My Account' section with 'Login' and 'Register' links, and an 'RSS Feeds' section.

archive:8080/xmlui/handle/600826/158

Figure 12: DSpace GUI



The screenshot shows the 'View Item' page for the 'RAI Test Video'. The breadcrumb trail is 'DSpace Home -> Presto4U Community -> D3.3 Test Collection -> View Item'. The page includes a 'Show simple item record' link and a table of metadata. The 'Files in this item' section shows a file named 'PrestoPRIME_RAI_0 ...' with a size of 223.7Mb and an unknown format. The 'This item appears in the following Collection(s)' section lists 'D3.3 Test Collection' with a sub-description 'AV files from Presto4U Dataset' and a 'Show simple item record' link. The right sidebar contains a 'Search DSpace' box, a 'Browse' section with links for 'All of DSpace' and 'This Collection', a 'My Account' section with 'Login' and 'Register' links, and an 'RSS Feeds' section.

dc.contributor.author	Gallo, Francesco	
dc.date.accessioned	2014-12-10T15:02:08Z	
dc.date.available	2014-12-10T15:02:08Z	
dc.date.issued	2014	
dc.identifier.uri	http://archive:8080/xmlui/handle/600826/160	
dc.language.iso	it	en_US
dc.publisher	RAI CRIT	en_US
dc.subject	Test video from Presto4U Dataset	en_US
dc.title	RAI Test Video	en_US
dc.type	Video	en_US

Figure 13: DSpace Repository

L describes how many functions have incorrect operation avoidance capabilities. This feature can be regarded as the degree to which the platform prevents the users from making mistakes, especially during the ingest process, that could affect the preservation of data. In particular it can be evaluated as:

$$L = (A+B+C+D+E)/5 = (1+1+1+1+1)/5 = 1.000.$$

Where A indicates whether there are required field to fill during the ingest process in order to clearly identify the data being ingested. B indicates if the platform checks the input formats to determine if they are compatible with its preservation capabilities (for instance the platform must be capable of migrating the format to another one). C indicates whether a check of the metadata is performed. D is the degree to which the user is guided through the ingestion process and E indicates if a check of the authenticity of the data is performed.

$$\text{Therefore } U_s = (K+L)/2 = 0.900.$$

Interpretation of test results: U_s value closer to 1 is better.

3.5.1.5 Reliability

Degree to which a system, product or component performs specified functions under specified conditions for a specified period of time.

Measurement function: $Re = (H+J+K+L)/4$ where

$Re =$ **Reliability**

H = Maturity

J = Availability

K = Fault Tolerance

L = Recoverability

As far as H is concerned, since the platforms taken into account are developed, supported and adopted by communities of users, this value should give a qualitative estimation of how wide the community behind the platform is and its degree of adoption. A score between 0 and 1 will be assigned. DSpace is one of the most adopted platforms and is supported by a wide community of users and developers thus the following score should be assigned:

$$H = 1.000.$$

J represents the availability of the platform. Since each of these systems is based upon web services, it is possible to assign a mark between 0 and 1 according to how the web services can be monitored by the user. Since the availability of web services can be tested

even from the command line using the tools provided with the default installation. It is reasonable to assign the following score:

$$J = 1.000.$$

K concerns how the platform can deal with user's errors or other failures without compromising the whole operation. It can be defined as:

$$K = (A+B+C)/3 = (1+0.5+1)/3 = 0.833.$$

Where A indicates if the platform allows to save a complete backup in order to restore the overall state of the platform itself in case of failure. B indicates the degree to which making a mistake affect the normal operability of the system. C indicates if the platform provides a validation mechanism for the ingestion process.

L indicates what is (the average) time the system takes to complete recovery from a failure. It is possible to take into account a given task, such as the ingestion process, and evaluate how the system reacts to the occurrence of a failure. In case the platform allows the user to cope with the failure and continue the ingestion the recoverability value should be close to 1. If, on the other hand, the platform requires the user to start the ingestion process from the beginning, this value should be close to 0. The user is not so clearly guided through the ingestion process as happens for other platforms so the following score should be assigned:

$$L = 0.500.$$

$$\text{Therefore } Re = (H+J+K+L)/4 = 0.833.$$

Interpretation of test results: Re value closer to 1 is better.

3.5.1.6 Security

Degree to which a product or system protects information and data so that persons or other products or systems have the degree of data access appropriate to their types and levels of authorization.

Measurement function: $Se = (H+J+K+L+M)/5$ where

Se = **Security**

H = Confidentiality

J = Integrity

K = Non-repudiation

L = Accountability

M = Authenticity

According to Section 2.1.6 of D3.2, H, J, K, L and M can be defined as follows:

H indicates how controllable is the access to the system. Since the platforms take advantage of web services to manage the ingested data, the security level provided by these web services is related to the degree of confidentiality. The technologies adopted for the deployment of a DSpace server (e.g. Apache Web Server and Apache Tomcat) leverage the best practices in securing web application and are supported by a huge community of developers for small to enterprise level installations. The assigned score should be:

H = 1.000.

J describes to what extent the system prevents unauthorised access to the data. This feature is closely related to the previous one so the security of the web services has to be taken into account. Since the access mechanism is safe enough to prevent unauthorized access it is possible to assign the following score:

J = 1.000.

K indicates what proportion of events requiring non-repudiation are processed. In order to satisfy this requirement the platform must be able to prove that an action has been performed so that it cannot be repudiated later. In case the system is provided with this capability it should get a high mark (from 0 to 1). In this case the following score should be assigned:

K = 1.000.

L describes how complete is the audit trail concerning the user access to the system and data. For the kind of systems being assessed, this feature may be related to the ACL capability so that the platform can assign a different access level to administrators with respect to users. The more complete is the set of rules that can be established, the higher is the score (between 0 and 1).

Since DSpace allows to specify several degree of accessibility the following score should be assigned:

$$L = 1.000.$$

M indicates how well does the system authenticate the identity of a subject or resource. It is implemented as:

$$M = A/B = 1.000.$$

Where A is the number of provided authentication methods (e.g., ID/password or IC card) and B is the total number of authentication methods specified in the requirements (e.g., ID/password or IC card).

$$\text{Therefore } Se = (H+J+K+L+M)/5 = 1.000.$$

Interpretation of test results: Se value closer to 1 is better.

3.5.1.7 Maintainability

Degree of effectiveness and efficiency with which a product or system can be modified by the intended maintainers.

Measurement function: $Ma = (H+K+L+M)/4$ where

Ma = Maintainability

H = Modularity

K = Analysability

L = Modifiability

M = Testability

In Section 2.1.7 of D3.2, H, K, L and M are described as follows.

H measures how strong is the relation between the components in a system or computer program. Certainly the platforms being assessed are made up of several components that have to interact with each other in order to make the system work properly. Considering the large communities of users and developers supporting these platforms, the interaction of the various components is granted by the maturity of the systems. Therefore an element that can be taken into account for the assessment is the possibility for the user to store data into a cloud storage. Keeping data separated from the system can be a benefit in

case of local failures. DSpace is built to be integrated and operate with a cloud storage, thus the score should be:

$$H = 1.000.$$

K indicates whether users can easily identify specific operation which caused failures. It is possible to consider the ingest process where the most part of errors may occur. In case the platform warns the user about failures and indicates the task that caused it, then the system should get a good mark (between 0 and 1). The user is warned in case the operation taken place fails so the following score should be assigned:

$$K = 1.000.$$

L indicates if the maintainer can easily modify the software to meet some modification requirement. An example of whether this requirement is satisfied is the possibility to switch from one database to another. This feature is related to the modularity. In the document no explicit reference to the possibility to migrate from one database to another thus the score should be:

$$L = 0.500.$$

M describes how completely are test functions and facilities implemented. It can be calculated as follows:

$$M = (A+B+C)/3 = (1+1+0)/3 = 0.667.$$

Where A is 1 in case the platform allows the user to perform dry run in order to verify the correctness of the operation, B is 1 if the platform provides diagnostic tools within its user interface and C is one in case it is possible to run a demo version of the platform in order to perform tests without compromising the actual data.

$$\text{Therefore } Ma = (H+K+L+M)/4 = 0.792.$$

Interpretation of test results: Ma value closer to 1 is better.

3.5.1.8 Portability

Degree of effectiveness and efficiency with which a system, product or component can be transferred from one hardware, software or other operational or usage environment to another.

Measurement function: $Po = (X+Y+Z)/3$ where

$Po = \text{Portability}$

$X = \text{Adaptability}$

$Y = \text{Installability}$

$Z = \text{Replaceability}$

The description of X, Y and Z is reported in Section 2.1.8 of D3.2. X indicates whether the software system is capable enough to adapt itself to different hardware environment. It is calculated as:

$$X = 1 - (A/B) = 1 - (0/3) = 1.000.$$

Where A is the number of operational functions of which tasks were not completed or not enough resulted to meet adequate levels during testing and B is the total number of functions which were tested in different hardware environment. The three functions used for the assessment are the ingestion of a content, an access through the user interface and the dissemination.

Y gives an idea of how much time and trouble is required to make an install. As far as this feature is concerned, the platform will be evaluated according to how clearly and completely is the installation process described in the documentation. The installation process explained but no virtual appliance are provided as happens for other platforms. Thus the score should be:

$$Y = 0.800.$$

Z measures the degree to which the system can be replaced by another one with the same purpose. The adoption of standard is a relevant element for the evaluation of this feature. Another element to take into account is whether is possible for the platform to be integrated with another one. A Dspace-based archive can be exported preserving the items, their relationship and the structure (e.g. the collections) into a format that can be processed by other preservation platforms. For example Archivemata can take as input the items exported in the DSpace format. Therefore the following score should be assigned:

$$Z = 1.000.$$

Therefore $Po = (X+Y+Z)/3 = 0.933$.

Interpretation of test results: Po value closer to 1 is better.

3.5.1.9 Summary of DSpace Assessment Results

The following table sums up the assessment results of DSpace:

Functional Suitability	$F_s =$	$(X+Z)/2 =$	0.879	X	0.872	X1	0.889	A	1
								B	9
						X2	0.900	C	1
						X3	0.750	D	10
								E	1
					Z	0.885	A	23	F
						B	26		
Performance Efficiency	$P_e =$	$Z =$	1.000	Z	1.000				
Compatibility	$C_o =$	$(X+Y)/2 =$	0.917	X	1.000	A	5	B	6
				Y	0.833				
Usability	$U_s =$	$(K+L)/2 =$	0.900	K	0.800				
				L	1.000	A	1		
						B	1		
						C	1		
						D	1		
						E	1		
Reliability	$R_e =$	$(H+J+K+L)/4 =$	0.833	H	1.000	A	1	B	0.5
				J	1.000				
				K	0.833				
				L	0.500				
						C	1		
Security	$S_e =$	$(H+J+K+L+M)/5 =$	1.000	H	1.000				
				J	1.000				
				K	1.000				
				L	1.000				
				M	1.000	A	2		
						B	2		

Maintainability	Ma =	(H+K+L+M)/4 =	0.792	H	1.000	A	1			
				K	1.000					
				L	0.500					
				M	0.667				B	1
									C	0
Portability	Po =	(X+Y+Z)/3 =	0.933	X	1.000	A	0			
						B	3			
				Y	0.800					
				Z	1.000					

Table 42: DSpace Assessment Summary

3.5.2 RODA

3.5.2.1 Functional Suitability

Degree to which a product or system provides functions that meet stated and implied needs when used under specified conditions.

Measurement function: $FS = (X+Z)/2$ where

FS = Functional Suitability

X = Functional Completeness

Z = Functional Appropriateness

X indicates how complete is the implementation according to requirement specifications.

As reported in Section 2.1.1 of D3.2, X can be calculated as follows:

$$X = (X1+X2*0.5+X3*0.25)/1.75 = (0.889+0.900*0.5+0.750*0.25)/1.75 = 0.872.$$

Where $X1=1-(1/9)$, $X2= 1-(1/10)$ and $X3= 1-(1/4)$. A is the number of missing or unsatisfying mandatory functions, B is the number of mandatory functions assessed in the evaluation, C is the number of missing or unsatisfying recommended functions assessed in the evaluation, D is the number of recommended functions, E is the number of missing or unsatisfying desirable functions assessed in the evaluation and F is the number of desirable functions. In particular, with respect to the table reported in Section 2, among the mandatory functions the content quality control (M9) was missing in the default installation. The functions R4 among the recommended and the D3 among the desirable ones were missing, too. Those features may be available as plug-ins or external tools.

Z describes how many functions with no problems are implemented for the appropriate functions for pursuing a specific task. Z can be calculated as:

$$Z = A/B = 23/26 = 0.885.$$

Where A is the sum of the scores of the implemented functions and B is the total amount of implemented functions.

$$\text{Therefore FS} = (X+Z)/2 = 0.879.$$

Interpretation of test results: FS value closer to 1 is better. The list of functions for the preservation platform considered during the assessment is presented in Section 2.5.1 of D3.2.

3.5.2.2 Performance Efficiency

Measurement function: $PE = Z$ where

PE = Performance Efficiency

Z = Capacity

Useful element for the evaluation of the capacity can be: the number of requests or simultaneous access per unit of time; the number of simultaneous jobs accepted in the ingestion queue or the number of tasks executed in parallel during a preservation workflow.

Such elements are strictly related to the hardware of the system into which the platforms are executed. For instance, since it is common for a new job or online request to throw a new thread, the availability of several computational units would improve the operation time of the platforms.

Due to the previous considerations, if the platform architecture allows a uniform distribution of the tasks, the capacity is scalable and thus the platform should get a good evaluation.

RODA supports the execution of multiple tasks in parallel, the hardware and the underlying technology determine the operational time of the system. The RODA community and KEEP SOLUTIONS support and maintain RODA to improve its performances. Finally, the building blocks of RODA, and the underlying Fedora data layer are maintained by wide communities and can be tuned in the most suitable way. For these reasons it seems reasonable to assign the following score:

$$Z = 1.000.$$

Therefore PE = 1.000.

Interpretation of test results: PE closer to 1 is better.

3.5.2.3 Compatibility

Measurement function: $Co = (X+Y)/2$ where

Co = **Compatibility**

X = Co-existence

Y = Interoperability

As explained in Section 2.1.3 of D3.2, X indicates how flexible is the product in sharing its environment with other products without adverse impacts on other products.

It is possible to evaluate if the platform requires an exclusive usage of a component such as the database. In case the database can be shared among other systems, the platform should get a good score for this feature (between 0 and 1).

In the documentation there are no explicit references to the possibility to make RODA operate with other platforms as happens with DSpace and Archivematica, but RODA is itself built on top of another platform, Fedora, which is adopted as digital repository in several projects. Therefore a reasonable score should be:

$X = 0.500$.

Y indicates how accurately is implementation of data exchange format determined between linking systems. It can be expressed as:

$Y = A/B = 0.500$.

Where A = 3 is the number of formats into which data can be exported in order to be exchanged with other platforms. B = 6 is the total number of data exportation formats provided by the platforms being assessed. The formats taken into account for this evaluation include among others PREMIS, DublinCore, METS and Simple Archive Format, just to name a few. For the list of supported format please refer to the documentation.

Therefore $Co = (X+Y)/2 = 0.500$.

Interpretation of test results: Co value closer to 1 is better.

3.5.2.4 Usability

Degree to which a product or system can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.

Measurement function: $Us = (K+L)/2$ where

$Us =$ **Usability**

$K =$ Operability

$L =$ User error protection

According to its definition, the operability indicates the degree to which the platform has attributes that make it easy to operate and control. A good estimation of K may come from the evaluation of the user interface provided by the platform. In case a clear and intuitive interface is provided the platform should get a good mark (between 0 and 1).

The GUI provided by RODA has been tested performing the most common preservation functions such as the ingest, the access and the archive administration. We tried typical producer and consumer tasks with the AV files available in the dataset and also tried administrative tasks such as archive monitoring. The RODA web GUI is supported by all popular browsers, looks quite clear and user friendly. After this test, we assign a relative score between the platforms under assessment. Concerning the GUI provided by RODA a reasonable score could be:

$K = 0.800$.

Figure 14: Example of RODA GUI

L describes how many functions have incorrect operation avoidance capabilities. This feature can be regarded as the degree to which the platform prevents the users from making mistakes, especially during the ingest process, that could affect the preservation of data. In particular it can be evaluated as:

$$L = (A+B+C+D+E)/5 = (1+1+0+1+1)/5 = 0.800.$$

Where A indicates whether there are required field to fill during the ingest process in order to clearly identify the data being ingested. B indicates if the platform checks the input formats to determine if they are compatible with its preservation capabilities (for instance the platform must be capable of migrating the format to another one). C indicates whether a check of the metadata is performed. D is the degree to which the user is guided through the ingestion process and E indicates if a check of the authenticity of the data is performed.

$$\text{Therefore } U_s = (K+L)/2 = 0.800.$$

Interpretation of test results: U_s value closer to 1 is better.

3.5.2.5 Reliability

Degree to which a system, product or component performs specified functions under specified conditions for a specified period of time.

Measurement function: $Re = (H+J+K+L)/4$ where

Re = Reliability

H = Maturity

J = Availability

K = Fault Tolerance

L = Recoverability

As far as H is concerned, since the platforms taken into account are developed, supported and adopted by communities of users, this value should give a qualitative estimation of how wide the community behind the platform is and its degree of adoption. A score between 0 and 1 will be assigned.

To the best of our knowledge, being a relatively new project, RODA is currently adopted by a reduced number of institution compared to large and long lasting projects such as DSpace and Archivematica. Nevertheless RODA is heavily OAI oriented thus could gain new adopters in the near future. Thus the following score should be assigned:

$$H = 0.500.$$

J represents the availability of the platform. Since each of these systems is based upon web services, it is possible to assign a mark between 0 and 1 according to how the web services can be monitored by the user. The web services have proved to be reliable enough to assign the following score:

$$J = 1.000.$$

K concerns how the platform can deal with user's errors or other failures without compromising the whole operation. It can be defined as:

$$K = (A+B+C)/3 = (1+1+0.5)/3 = 0.833.$$

Where A indicates if the platform allows to save a complete backup in order to restore the overall state of the platform itself in case of failure. B indicates the degree to which making

a mistake affect the normal operability of the system. C indicates if the platform provides a validation mechanism for the ingestion process.

L indicates what is (the average) time the system takes to complete recovery from a failure. It is possible to take into account a given task, such as the ingestion process, and evaluate how the system reacts to the occurrence of a failure. In case the platform allows the user to cope with the failure and continue the ingestion the recoverability value should be close to 1. If, on the other hand, the platform requires the user to start the ingestion process from the beginning, this value should be close to 0. The user is clearly guided through the ingestion process. The process is not divided in many steps as happens for Archivemantica resulting in a lower failure tolerance. Thus the following score should be assigned:

$$L = 0.500.$$

$$\text{Therefore } Re = (H+J+K+L)/4 = 0.708.$$

Interpretation of test results: Re value closer to 1 is better.

3.5.2.6 Security

Degree to which a product or system protects information and data so that persons or other products or systems have the degree of data access appropriate to their types and levels of authorization.

Measurement function: $Se = (H+J+K+L+M)/5$ where

Se = **Security**

H = Confidentiality

J = Integrity

K = Non-repudiation

L = Accountability

M = Authenticity

According to Section 2.1.6 of D3.2, H, J, K, L and M can be defined as follows:

H indicates how controllable is the access to the system. Since the platforms take advantage of web services to manage the ingested data, the security level provided by these web services is related to the degree of confidentiality.

The technologies adopted for the deployment of a RODA (Fedora) server (e.g. Apache Web Server and Apache Tomcat) leverage the best practices in securing web application and are supported by a huge community of developers for small to enterprise level installations. The assigned score should be:

$$H = 1.000.$$

J describes to what extent the system prevents unauthorised access to the data. This feature is closely related to the previous one so the security of the web services has to be taken into account. Since the access mechanism is safe enough to prevent unauthorized access it is possible to assign the following score:

$$J = 1.000.$$

K indicates what proportion of events requiring non-repudiation are processed. In order to satisfy this requirement the platform must be able to prove that an action has been performed so that it cannot be repudiated later. In case the system is provided with this capability it should get a high mark (from 0 to 1). In this case the following score should be assigned:

$$K = 0.500.$$

L describes how complete is the audit trail concerning the user access to the system and data. For the kind of systems being assessed, this feature may be related to the ACL capability so that the platform can assign a different access level to administrators with respect to users. The more complete is the set of rules that can be established, the higher is the score (between 0 and 1). Since RODA allows to specify several degree of accessibility the following score should be assigned:

$$L = 1.000.$$

M indicates how well does the system authenticate the identity of a subject or resource. It is implemented as:

$$M = A/B = 1.000.$$

Where A is the number of provided authentication methods (e.g., ID/password or IC card) and B is the total number of authentication methods specified in the requirements (e.g., ID/password or IC card).

Therefore $Se = (H+J+K+L+M)/5 = 0.900$.

Interpretation of test results: Se value closer to 1 is better.

3.5.2.7 Maintainability

Degree of effectiveness and efficiency with which a product or system can be modified by the intended maintainers.

Measurement function: $Ma = (H+K+L+M)/4$ where

Ma = Maintainability

H = Modularity

K = Analysability

L = Modifiability

M = Testability

In Section 2.1.7 of D3.2, H, K, L and M are described as follows.

H measures how strong is the relation between the components in a system or computer program. Certainly the platforms being assessed are made up of several components that have to interact with each other in order to make the system work properly. Considering the large communities of users and developers supporting these platforms, the interaction of the various components is granted by the maturity of the systems. Therefore an element that can be taken into account for the assessment is the possibility for the user to store data into a cloud storage. Keeping data separated from the system can be a benefit in case of local failures.

RODA is built on top of Fedora which acts as a data storage layer. Fedora supports several storage configurations including cloud (see for example the integration with DuraCloud, the cloud project developed by DuraSpace, the community supporting Fedora and DSpace). Therefore a reasonable score should be:

H = 1.000.

K indicates whether users can easily identify specific operation which caused failures. It is possible to consider the ingest process where the most part of errors may occur. In case the platform warns the user about failures and indicates the task that caused it, then the system should get a good mark (between 0 and 1). The user is clearly warned in case the operation taken place fails so the following score should be assigned:

$$K = 1.000.$$

L indicates if the maintainer can easily modify the software to meet some modification requirement. An example of whether this requirement is satisfied is the possibility to switch from one database to another. This feature is related to the modularity. In the document no explicit reference to the possibility to switch from one database to another thus the score should be:

$$L = 0.500.$$

M describes how completely are test functions and facilities implemented. It can be calculated as follows:

$$M = (A+B+C)/3 = (0+1+1)/3 = 0.667.$$

Where A is 1 in case the platform allows the user to perform dry run in order to verify the correctness of the operation, B is 1 if the platform provides diagnostic tools within its user interface and C is one in case it is possible to run a demo version of the platform in order to perform tests without compromising the actual data.

$$\text{Therefore } Ma = (H+K+L+M)/4 = 0.792.$$

Interpretation of test results: Ma value closer to 1 is better.

3.5.2.8 Portability

Degree of effectiveness and efficiency with which a system, product or component can be transferred from one hardware, software or other operational or usage environment to another.

Measurement function: $Po = (X+Y+Z)/3$ where

$Po = \text{Portability}$

$X = \text{Adaptability}$

$Y = \text{Installability}$

$Z = \text{Replaceability}$

The description of X, Y and Z is reported in Section 2.1.8 of D3.2. X indicates whether the software system is capable enough to adapt itself to different hardware environment. It is calculated as:

$$X = 1 - (A/B) = 1 - (0/3) = 1.000.$$

Where A is the number of operational functions of which tasks were not completed or not enough resulted to meet adequate levels during testing and B is the total number of functions which were tested in different hardware environment. The three functions used for the assessment are the ingestion of a content, an access through the user interface and the dissemination.

Y gives an idea of how much time and trouble is required to make an install. As far as this feature is concerned, the platform will be evaluated according to how clearly and completely is the installation process described in the documentation. The installation process explained. Moreover a useful demo version is provided. Thus the score should be:

$$Y = 1.000.$$

Z measures the degree to which the system can be replaced by another one with the same purpose. The adoption of standard is a relevant element for the evaluation of this feature. Another element to take into account is whether is possible for the platform to be integrated with another one.

RODA can interoperate with other platforms through the data storage layer since Fedora can interact with other preservation systems, such as DSpace by means of data model mapping. Therefore the following score should be assigned:

$$Z = 1.000.$$

$$\text{Therefore } P_o = (X+Y+Z)/3 = 1.000.$$

Interpretation of test results: P_o value closer to 1 is better.

3.5.2.9 Summary of RODA Assessment Results

The following table sums up the assessment results of RODA:

Functional Suitability	$F_s =$	$(X+Z)/2 =$	0.879	X	0.872	X1	0.889	A	1
								B	9
						X2	0.900	C	1

				Z	0.885	A	23	D	10
						B	26	E	1
								F	4
Performance Efficiency	Pe =	Z =	1.000	Z	1.000				
Compatibility	Co =	(X+Y)/2 =	0.500	X	0.500				
				Y	0.500	A	3		
						B	6		
Usability	Us =	(K+L)/2 =	0.800	K	0.800				
				L	0.800	A	1		
						B	1		
						C	0		
						D	1		
						E	1		
Reliability	Re =	(H+J+K+L)/4 =	0.708	H	0.500				
				J	1.000				
				K	0.833	A	1		
						B	1		
						C	0.5		
				L	0.500				
Security	Se =	(H+J+K+L+M)/5 =	0.900	H	1.000				
				J	1.000				
				K	0.500				
				L	1.000				
				M	1.000	A	2		
						B	2		
Maintainability	Ma =	(H+K+L+M)/4 =	0.792	H	1.000				
				K	1.000				
				L	0.500				
				M	0.667	A	0		
						B	1		
						C	1		
Portability	Po =	(X+Y+Z)/3 =	1.000	X	1.000	A	0		
						B	3		
				Y	1.000				
				Z	1.000				

Table 43: Summary of RODA results

3.5.3 Archivematica

In this section we describe the assessment of Archivematica which was also included in year one evaluation. The motivation behind the reassessment is twofold: on one hand the new releases of Archivematica provided several improvements especially concerning the installation process (from pre-built virtual machines to supported packages in Ubuntu repository); on the other hand Archivematica is under investigation by several CoP in the project (see WP2 deliverables).

3.5.3.1 Functional Suitability

Degree to which a product or system provides functions that meet stated and implied needs when used under specified conditions.

Measurement function: $FS = (X+Z)/2$ where

FS = Functional Suitability

X = Functional Completeness

Z = Functional Appropriateness

X indicates how complete is the implementation according to requirement specifications.

As reported in Section 2.1.1 of D3.2, X can be calculated as follows:

$$X = (X1+X2*0.5+X3*0.25)/1.75 = (0.889+0.778*0.5+0.333*0.25)/1.75 = 0.778.$$

Where $X1=1-(1/9)$, $X2= 1-(2/9)$ and $X3= 1-(2/3)$. A is the number of missing or unsatisfying mandatory functions, B is the number of mandatory functions assessed in the evaluation, C is the number of missing or unsatisfying recommended functions assessed in the evaluation, D is the number of recommended functions, E is the number of missing or unsatisfying desirable functions assessed in the evaluation and F is the number of desirable functions. In particular, with respect to the table reported in Section 2, among the mandatory functions the content quality control (M9) was missing in the default installation. The functions R4 and R5 among the recommended and the D2 and D3 among the desirable ones were missing, too. Those features may be available as plug-ins or external tools. To the best of our knowledge, the customization of the platform with additional add-ons or plug-ins is not supported out of the box and no plug-in mechanism is available, compared to the other evaluated platforms. Hence the integration of other systems apparently requires the user to modify the source code and recompile the whole platform.

Z describes how many functions with no problems are implemented for the appropriate functions for pursuing a specific task. Z can be calculated as:

$$Z = A/B = 21/26 = 0.808$$

Where A is the sum of the scores of the implemented functions and B is the total amount of implemented functions.

$$\text{Therefore FS} = (X+Z)/2 = 0.793.$$

Interpretation of test results: FS value closer to 1 is better. The list of functions for the preservation platform considered during the assessment is presented in Section 2.5.1 of D3.2.

3.5.3.2 Performance Efficiency

Measurement function: $PE = Z$ where

PE = Performance Efficiency

Z = Capacity

Useful element for the evaluation of the capacity can be: the number of requests or simultaneous access per unit of time; the number of simultaneous jobs accepted in the ingestion queue or the number of tasks executed in parallel during a preservation workflow.

Such elements are strictly related to the hardware of the system into which the platforms are executed. For instance, since it is common for a new job or online request to throw a new thread, the availability of several computational units would improve the operation time of the platforms.

Due to the previous considerations, if the platform architecture allows a uniform distribution of the tasks, the capacity is scalable and thus the platform should get a good evaluation.

Archivematica is developed using a Python-based Django MVC framework and implements a micro-services pattern. Micro-services can be distributed to processing clusters for highly scalable configurations. Archivematica can be installed either in a virtualization environment or directly on dedicated hardware via its own Ubuntu repository. Archivematica can be deployed in multi-node, distributed processing configuration to support large-scale, resource-intensive production environments, where a significant number of concurrent tasks may be required. The Archivematica community supporting and maintaining Archivematica continuously improves the software to fix memory leaks and other issues affecting Archivematica performances. For these reasons it seems reasonable to assign the following score:

$$Z = 1.000.$$

Therefore PE = 1.000.

Interpretation of test results: PE closer to 1 is better.

3.5.3.3 Compatibility

Measurement function: $Co = (X+Y)/2$ where

$Co =$ **Compatibility**

$X =$ Co-existence

$Y =$ Interoperability

As explained in Section 2.1.3 of D3.2, X indicates how flexible is the product in sharing its environment with other products without adverse impacts on other products.

It is possible to evaluate if the platform requires an exclusive usage of a component such as the database. In case the database can be shared among other systems, the platform should get a good score for this feature (between 0 and 1). Archivemática can both be used through a virtual appliance or installed in a software environment. It can also operate with DSpace so the score should be:

$X = 1.000.$

Y indicates how accurately is implementation of data exchange format determined between linking systems. It can be expressed as:

$Y = A/B = 0.500.$

Where $A = 3$ is the number of formats into which data can be exported in order to be exchanged with other platforms. $B = 6$ is the total number of data exportation formats provided by the platforms being assessed.

Therefore $Co = (X+Y)/2 = 0.750.$

Interpretation of test results: Co value closer to 1 is better.

3.5.3.4 Usability

Degree to which a product or system can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.

Measurement function: $Us = (K+L)/2$ where

$Us =$ **Usability**

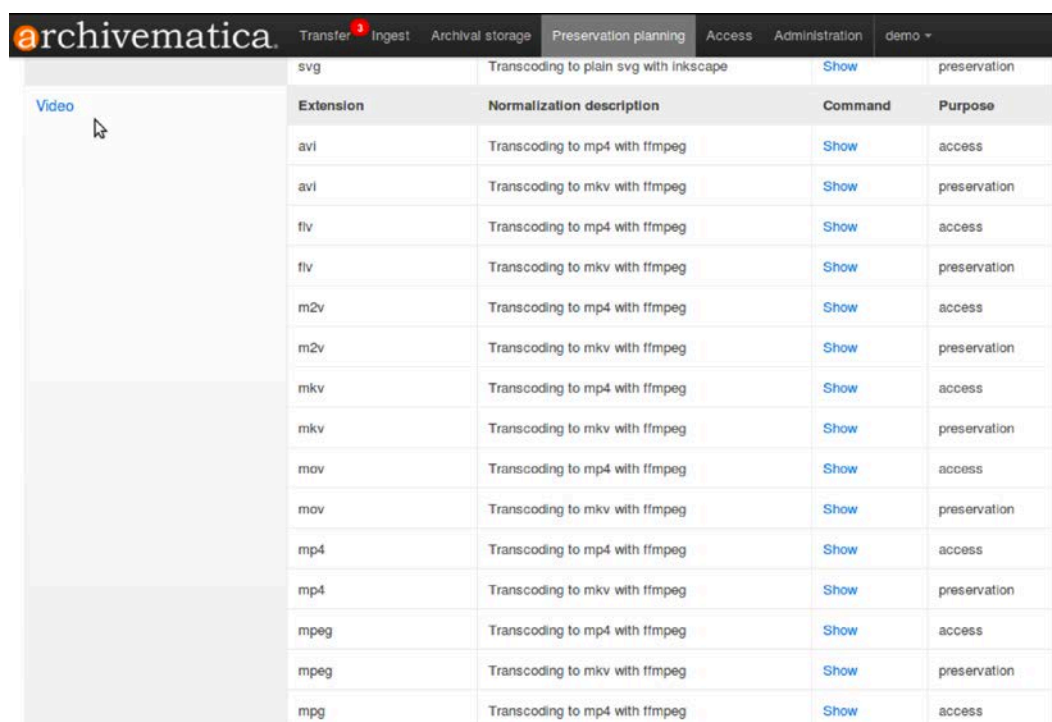
$K =$ Operability

$L =$ User error protection

According to its definition, the operability indicates the degree to which the platform has attributes that make it easy to operate and control. A good estimation of K may come from the evaluation of the user interface provided by the platform. In case a clear and intuitive interface is provided the platform should get a good mark (between 0 and 1).

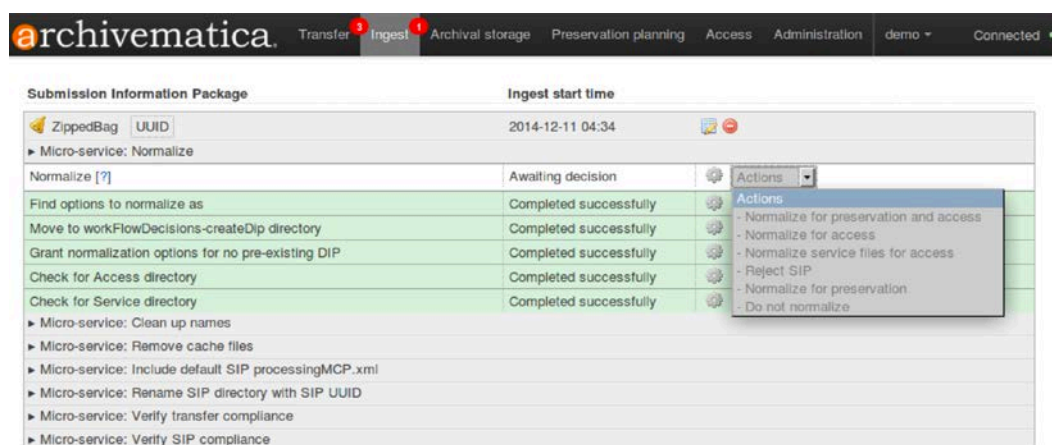
We tested the GUI provided by Archivemata when performing the most common preservation functions such as the ingest, the access and the archive administration. We tried typical producer and consumer tasks with the AV files available in the dataset and also tried administrative tasks such as archive monitoring. The new dashboard available with last releases includes advanced features, enabling users to process, monitor and control the workflows. The dashboard provides also a monitoring interface collecting the status of system events and it can be used to easily control and trigger specific micro-services, edit archived content and provide preservation planning information. Concerning administration tasks the user can, for example, manage storage locations, configure the micro-services of the ingest workflow, modify the preservation plans and manage users and ACLs. Further improvements of the dashboard have already been announced for the incoming releases. Since the GUI provided by Archivemata was clear and intuitive, a reasonable score could be:

$K = 1.000$.



Extension	Normalization description	Command	Purpose
svg	Transcoding to plain svg with inkscape	Show	preservation
avi	Transcoding to mp4 with ffmpeg	Show	access
avi	Transcoding to mkv with ffmpeg	Show	preservation
flv	Transcoding to mp4 with ffmpeg	Show	access
flv	Transcoding to mkv with ffmpeg	Show	preservation
m2v	Transcoding to mp4 with ffmpeg	Show	access
m2v	Transcoding to mkv with ffmpeg	Show	preservation
mkv	Transcoding to mp4 with ffmpeg	Show	access
mkv	Transcoding to mkv with ffmpeg	Show	preservation
mov	Transcoding to mp4 with ffmpeg	Show	access
mov	Transcoding to mkv with ffmpeg	Show	preservation
mp4	Transcoding to mp4 with ffmpeg	Show	access
mp4	Transcoding to mkv with ffmpeg	Show	preservation
mpeg	Transcoding to mp4 with ffmpeg	Show	access
mpeg	Transcoding to mkv with ffmpeg	Show	preservation
mpg	Transcoding to mp4 with ffmpeg	Show	access

Figure 15: Normalisation formats for videos



Submission Information Package	Ingest start time
ZippedBag UUID	2014-12-11 04:34
Micro-service: Normalize	Awaiting decision
Normalize [?]	Completed successfully
Find options to normalize as	Completed successfully
Move to workFlowDecisions-createDip directory	Completed successfully
Grant normalization options for no pre-existing DIP	Completed successfully
Check for Access directory	Completed successfully
Check for Service directory	Completed successfully
Micro-service: Clean up names	
Micro-service: Remove cache files	
Micro-service: Include default SIP processingMCP.xml	
Micro-service: Rename SIP directory with SIP UUID	
Micro-service: Verify transfer compliance	
Micro-service: Verify SIP compliance	

Figure 16: Ingest workflow - Progress monitoring and management

As shown in the figure above, Archivematica defines normalization formats for all content types and for each input content assigns a preservation format and an access format. The tools used for content transformation are embedded in the platform and are managed using micro-services.

The picture above shows an example of how the management of micro-services takes place by means of Archivematica dashboard. For example each step in the ingest can be checked, re-executed and customized (using available configuration) by the user.

L describes how many functions have incorrect operation avoidance capabilities. This feature can be regarded as the degree to which the platform prevents the users from

making mistakes, especially during the ingest process, that could affect the preservation of data. In particular it can be evaluated as:

$$L = (A+B+C+D+E)/5 = (1+1+1+1+1)/5 = 1.000.$$

Where A indicates whether there are required field to fill during the ingest process in order to clearly identify the data being ingested. B indicates if the platform checks the input formats to determine if they are compatible with its preservation capabilities (for instance the platform must be capable of migrating the format to another one). C indicates whether a check of the metadata is performed. D is the degree to which the user is guided through the ingestion process and E indicates if a check of the authenticity of the data is performed.

$$\text{Therefore } U_s = (K+L)/2 = 1.000.$$

Interpretation of test results: U_s value closer to 1 is better.

3.5.3.5 Reliability

Degree to which a system, product or component performs specified functions under specified conditions for a specified period of time.

Measurement function: $Re = (H+J+K+L)/4$ where

$Re =$ **Reliability**

H = Maturity

J = Availability

K = Fault Tolerance

L = Recoverability

As far as H is concerned, since the platforms taken into account are developed, supported and adopted by communities of users, this value should give a qualitative estimation of how wide the community behind the platform is and its degree of adoption. A score between 0 and 1 will be assigned. Archivematica is a widely adopted platform and is supported by a large community of users and developers thus the following score should be assigned:

$$H = 1.000.$$

J represents the availability of the platform. Since each of these systems is based upon web services, it is possible to assign a mark between 0 and 1 according to how the web services can be monitored by the user. The web services have proved to be reliable enough to assign the following score:

$$J = 1.000.$$

K concerns how the platform can deal with user's errors or other failures without compromising the whole operation. It can be defined as:

$$K = (A+B+C)/3 = (0.5+1+1)/3 = 0.833.$$

Where A indicates if the platform allows to save a complete backup in order to restore the overall state of the platform itself in case of failure. B indicates the degree to which making a mistake affect the normal operability of the system. C indicates if the platform provides a validation mechanism for the ingestion process.

L indicates what is (the average) time the system takes to complete recovery from a failure. It is possible to take into account a given task, such as the ingestion process, and evaluate how the system reacts to the occurrence of a failure. In case the platform allows the user to cope with the failure and continue the ingestion the recoverability value should be close to 1. If, on the other hand, the platform requires the user to start the ingestion process from the beginning, this value should be close to 0. The user is clearly guided through the ingestion process and in case of failure it is possible to repeat a specific step rather than restart from the beginning. Thus the following score should be assigned:

$$L = 1.000.$$

$$\text{Therefore } Re = (H+J+K+L)/4 = 0.958.$$

Interpretation of test results: Re value closer to 1 is better.

3.5.3.6 Security

Degree to which a product or system protects information and data so that persons or other products or systems have the degree of data access appropriate to their types and levels of authorization.

Measurement function: $Se = (H+J+K+L+M)/5$ where

$$Se = \text{Security}$$

H = Confidentiality

J = Integrity

K = Non-repudiation

L = Accountability

M = Authenticity

According to Section 2.1.6 of D3.2, H, J, K, L and M can be defined as follows:

H indicates how controllable is the access to the system. Since the platforms take advantage of web services to manage the ingested data, the security level provided by these web services is related to the degree of confidentiality.

The technologies adopted for the deployment of a Archivemata server (e.g. Python and Django MVC framework) leverage the best practices in securing web applications and are supported by a huge community of developers for small to enterprise level installations. Archivemata uses an encryption algorithm to secure password and other confidential information in combination with Django security mechanisms. The assigned score should be:

H = 1.000.

J describes to what extent the system prevents unauthorised access to the data. This feature is closely related to the previous one so the security of the web services has to be taken into account. Since the access mechanism is safe enough to prevent unauthorized access it is possible to assign the following score:

J = 1.000.

K indicates what proportion of events requiring non-repudiation are processed. In order to satisfy this requirement the platform must be able to prove that an action has been performed so that it cannot be repudiated later. In case the system is provided with this capability it should get a high mark (from 0 to 1). In this case the following score should be assigned:

K = 1.000.

L describes how complete is the audit trail concerning the user access to the system and data. For the kind of systems being assessed, this feature may be related to the ACL capability so that the platform can assign a different access level to administrators with respect to users. The more complete is the set of rules that can be established, the higher

is the score (between 0 and 1). Since Archivemataca allows to specify several degree of accessibility the following score should be assigned:

$$L = 1.000.$$

M indicates how well does the system authenticate the identity of a subject or resource. It is implemented as:

$$M = A/B = 1.000.$$

Where A is the number of provided authentication methods (e.g., ID/password or IC card) and B is the total number of authentication methods specified in the requirements (e.g., ID/password or IC card).

$$\text{Therefore } Se = (H+J+K+L+M)/5 = 1.000.$$

Interpretation of test results: Se value closer to 1 is better.

3.5.3.7 Maintainability

Degree of effectiveness and efficiency with which a product or system can be modified by the intended maintainers.

Measurement function: $Ma = (H+K+L+M)/4$ where

Ma = Maintainability

H = Modularity

K = Analysability

L = Modifiability

M = Testability

In Section 2.1.7 of D3.2, H, K, L and M are described as follows.

H measures how strong is the relation between the components in a system or computer program. Certainly the platforms being assessed are made up of several components that have to interact with each other in order to make the system work properly. Considering the large communities of users and developers supporting these platforms, the interaction of the various components is granted by the maturity of the systems. Therefore an element

that can be taken into account for the assessment is the possibility for the user to store data into a cloud storage. Keeping data separated from the system can be a benefit in case of local failures. Since it is possible to run Archivematica in a cloud environment and also to export archived content using web technologies such as AtoM and CONTENTdm the score should be:

$$H = 1.000.$$

K indicates whether users can easily identify specific operation which caused failures. It is possible to consider the ingest process where the most part of errors may occur. In case the platform warns the user about failures and indicates the task that caused it, then the system should get a good mark (between 0 and 1). The user is clearly warned in case the operation taken place fails so the following score should be assigned:

$$K = 1.000.$$

L indicates if the maintainer can easily modify the software to meet some modification requirement. An example of whether this requirement is satisfied is the possibility to switch from one database to another. This feature is related to the modularity. Archivematica stores data using the filesystem, ElasticSearch (long term) and a MySQL database (short term processing). In the documentation there is no explicit reference to the possibility to switch from one database to another thus the score should be:

$$L = 0.500.$$

M describes how completely are test functions and facilities implemented. It can be calculated as follows:

$$M = (A+B+C)/3 = (0+1+1)/3 = 0.667.$$

Where A is 1 in case the platform allows the user to perform dry run in order to verify the correctness of the operation, B is 1 if the platform provides diagnostic tools within its user interface and C is one in case it is possible to run a demo version of the platform in order to perform tests without compromising the actual data.

$$\text{Therefore } Ma = (H+K+L+M)/4 = 0.792.$$

Interpretation of test results: Ma value closer to 1 is better.

3.5.3.8 Portability

Degree of effectiveness and efficiency with which a system, product or component can be transferred from one hardware, software or other operational or usage environment to another.

Measurement function: $Po = (X+Y+Z)/3$ where

$Po =$ **Portability**

$X =$ Adaptability

$Y =$ Installability

$Z =$ Replaceability

The description of X, Y and Z is reported in Section 2.1.8 of D3.2. The evaluation of these features may differ from the one described in Section 2.1.8 in order to better adapt to the assessment of digital platforms.

X indicates whether the software system is capable enough to adapt itself to different hardware environment. It is calculated as:

$$X = 1 - (A/B) = 1 - (0/3) = 1.000.$$

Where A is the number of operational functions of which tasks were not completed or not enough resulted to meet adequate levels during testing and B is the total number of functions which were tested in different hardware environment. The three functions used for the assessment are the ingestion of a content, an access through the user interface and the dissemination.

Y gives an idea of how much time and trouble is required to make an install. As far as this feature is concerned, the platform will be evaluated according to how clearly and completely is the installation process described in the documentation. The installation process explained. Moreover a virtual appliance is provided. Thus the score should be:

$$Y = 1.000.$$

Z measures the degree to which the system can be replaced by another one with the same purpose. The adoption of standard is a relevant element for the evaluation of this feature. Another element to take into account is whether is possible for the platform to be integrated with another one. Archivematica can be replaced with other platforms so the following score should be assigned:

$$Z = 1.000.$$

Therefore $P_o = (X+Y+Z)/3 = 1.000$.

Interpretation of test results: P_o value closer to 1 is better.

3.5.3.9 Summary of Archivematica Assessment Results

The following table sums up the assessment results of Archivematica:

Functional Suitability	$F_s =$	$(X+Z)/2 =$	0.793	X	0.778	X1	0.889	A	1
						X2	0.778	B	9
						X3	0.333	C	2
								D	9
								E	2
								F	3
				Z	0.808	A	21		
						B	26		
Performance Efficiency	$P_e =$	$Z =$	1.000	Z	1.000				
Compatibility	$C_o =$	$(X+Y)/2 =$	0.750	X	1.000				
				Y	0.500	A	3		
						B	6		
Usability	$U_s =$	$(K+L)/2 =$	1.000	K	1.000				
				L	1.000	A	1		
						B	1		
						C	1		
						D	1		
						E	1		
Reliability	$R_e =$	$(H+J+K+L)/4 =$	0.958	H	1.000				
				J	1.000				
				K	0.833	A	0.5		
						B	1		
						C	1		
				L	1.000				
Security	$S_e =$	$(H+J+K+L+M)/5 =$	1.000	H	1.000				
				J	1.000				
				K	1.000				
				L	1.000				

				M	1.000	A	2				
						B	2				
Maintainability	Ma =	$(H+K+L+M)/4 =$	0.792	H	1.000						
				K	1.000						
				L	0.500						
				M	0.667	A	0				
						B	1				
						C	1				
Portability	Po =	$(X+Y+Z)/3 =$	1.000	X	1.000	A	0				
						B	3				
				Y	1.000						
				Z	1.000						

Table 44: Archivematica summary of results

4 Final Presto4U Dataset

In addition to the dataset described for Year 1 (see D3.2 [4]), we have the following additions. In order to run storage tests chose the following existing and free to download datasets.

- A dataset with a wide range of file sizes, e.g. ICoSOLE with a wide variety of file sizes.
- A dataset with some realistic cases, e.g. BLIP1000 which includes directory structures with data and metadata.

Both datasets provide sufficient licensing (Creative Commons) to be used for testing. They provide real AV data, including audio, video, still images as well as relevant metadata. ICoSOLE and BLIP10000 can provide the necessary data to create realistic tests for storage systems.

4.1 ICoSOLE

Immersive Coverage of Spatially Outspread Live Events is an FP7 project¹⁴ that aims at developing a platform that enables users to experience live events which are spatially spread out. The project will develop a platform for a context-adapted hybrid broadcast-Internet service, providing efficient tools for capture, production and distribution of audio-visual content captured by a heterogeneous set of devices spread over the event site.

The dataset is accessible under the Creative Commons (CC BY-NC 4.0) license. It consists of a large number of video recordings, 2119 files, of different file sizes that range from 2MB to 120GB files. This makes the ICoSOLE dataset a good candidate for simple file tests.

The following figures show the distribution of file number and sizes across the whole range of the dataset. The majority of the files, i.e. 91% are less than 64MB and represent 11.8% of data in the dataset. Files in the range of 16GB-32GB represent 40% of data in the dataset.

¹⁴ <http://icosole.eu/>

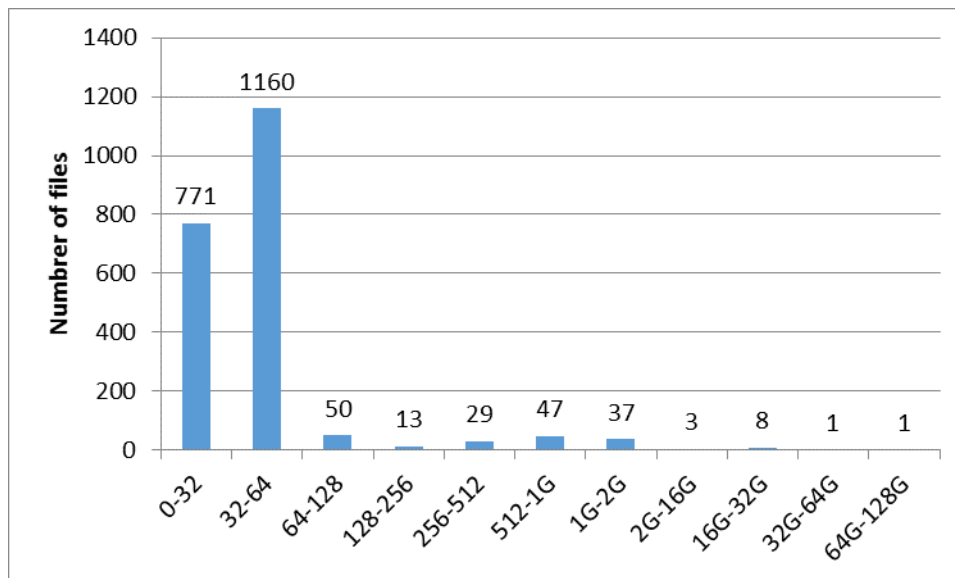


Figure 4-1 ICOSOLE dataset number of files distribution

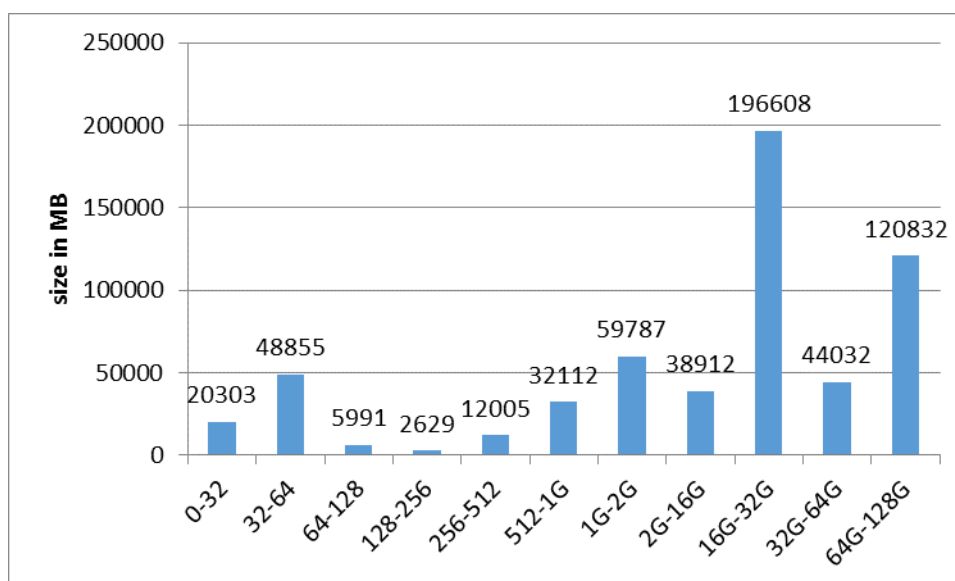


Figure 4-2 ICOSOLE dataset file size distribution

4.2 BLIP10000 Dataset

The Blip10000 dataset created by the PetaMedia NoE¹⁵ contains 14,838 Creative Commons videos from blip.tv, and corresponding user provided meta-data. The data content is generally referred to as semi-professional user generated (SPUG) content. The data comprises a total of 3,288 hours of data and is divided into development and test sets, of 5,288 and 9,550 videos respectively. The datasets include a combination of information from audio-visual content, user-contributed metadata, automatic speech recognition (ASR) transcripts, and social networks.

A large number of datasets in BLIP10K are included in a single or multipart compressed files, which range up to 120GB. The majority of video files are not very large, typically less than 1 GB. The following figure shows the directory structure of the Blip10K dataset.

¹⁵ <http://www.petamedia.eu/>




























Folder	Usage	Size	Contents
▼ mmsys		100 % 1.2 TB	3 items
▼ 2013		80.4 % 983.9 GB	11 items
▼ blip		95.1 % 935.8 GB	20 items
▼ Test		64.2 % 600.4 GB	9 items
Video		98.9 % 593.6 GB	9509 items
Shots		0.7 % 4.5 GB	8 items
AlternativeVideo		0.4 % 2.3 GB	68 items
Metadata		0.0 % 6.5 MB	2 items
GroundTruth		0.0 % 832.5 kB	4 items
▼ Dev		32.1 % 300.5 GB	9 items
Video		98.5 % 296.0 GB	5255 items
Shots		1.1 % 3.3 GB	7 items
AlternativeVideo		0.4 % 1.1 GB	33 items
Metadata		0.0 % 4.0 MB	2 items
GroundTruth		0.0 % 295.9 kB	3 items
ASR		3.6 % 33.4 GB	18 items
SocialData		0.2 % 1.6 GB	4 items
▶ sopcast		2.4 % 23.9 GB	10 items
multimodalmusic		1.3 % 12.9 GB	2 items
social2012		0.6 % 6.0 GB	8 items
fashion		0.2 % 2.4 GB	2 items
markedhead		0.2 % 1.9 GB	2 items
androidvideo		0.1 % 903.6 MB	5 items
▶ ddash		0.0 % 43.3 MB	13 items
▶ pathbandwidth		0.0 % 10.0 MB	14 items
jiku		0.0 % 0 bytes	1 item
2014		19.6 % 239.5 GB	7 items

Figure 3 Blip10K dataset directory structure

4.3 Dataset for MXF tool chain testing

The audiovisual material that has been used for the tests includes all the files provided within the P4U project plus a certain amount of additional files that has been judged necessary for having a sufficient quantity and diversity of file formats. In particular were added some MPEG transport stream files (including MPEG2 and H264 essence) and some MXF files containing uncompressed essence.

The detailed composition of the dataset is shown in Table 45, differentiated by typology.

File Type	Number of file analyzed	Nbr of video streams	Nbr of audio streams	MB	Total duration
MXF-D10	8	8	8	5930	00:18:00
MXF-XDCAM	12	12	12*8 = 96	9102	00:20:02

MXF-Proxy	12	12	$12 \times 4 = 48$	315	00:20:02
MXF-Uncompressed ¹⁶	11	11	$12 \times 4 = 48$	15280	00:09:24
MP4	9	9	9	25168	11:02:57
MP4-H264 Proxy	20	20	20	1006	01:39:40
TS - MPEG2-SD, AVC-HD ¹⁷	18	$18 \times 3 = 54$ (36 MPEG2 and 18 H.264)	$18 \times 5 = 90$ (72 MP2 and 18 AC-3)	94237	11:05:56
MOV – Prores	4	4	4	10918	00:11:57
OGV - flv	20	20	20	888	03:13:32
	114	150	343	162844	28:21:30

Table 45 - Composition of the data set

¹⁶ Not in the shared P4U dataset

¹⁷ Not in the shared P4U dataset

5 Conclusion

In this report we have presented the work done in year 2 with regards to the research output assessment exercise carried out as defined in WP3 Task 3.2 'Preservation Research Technology Watch and Assessment'. Both quantitative and qualitative analysis of the tools was performed against a series of criteria and metrics which measured characteristics such as functionality, ease of installation, robustness, performance, and scalability. During the course of the second year, we have also updated and finalised the Presto4U dataset used for this assessment task. The purpose of this dataset is twofold. Firstly to act as a test dataset for the assessment of research outputs carried out internally within the project, and secondly the aim make it available (after agreeing licencing terms and conditions) at the end of the project which can be used for testing tools outside the project in the AV preservation domain.

In the first part of the deliverable, we focussed on the new tools selected for assessment in year 2 along with the updates made to assessment templates. This meant some of the tools (e.g. Archivematica) needed to be re-evaluated based on new criteria and measurement functions. Next, we presented the detailed evaluation results of the RO Year 2 assessment. This includes the addition of two new categories of tools – vocabulary mapping and technical metadata extractors for which new assessment templates had to be defined. Finally, we presented the final Presto4U dataset. In our experience many of the open source and free to download datasets are sufficient and provide good coverage in terms of variation of file formats and file types for the testing of storage and preservation tools. However, specialised datasets had to be created for testing the MXF toolchain. The final dataset is hence a combination of existing and in-house generated files. One highlight of the assessment task was that we were able to engage with major commercial partners (Sony and Front Porch Digital) and performed a detailed assessment of storage based hardware solution for archiving. The results of these tests (LTO6 and Optical Drive) will be presented as a separate document and will remain confidential until it is approved by the commercial entities for public consumption.

In terms of future work, we need to look into the aspect of using the results of these quantitative and qualitative evaluations as part of brokerage in WP4. Also, the licence terms and conditions for the release of the Presto4U dataset need to be agreed upon. Deliverables D3.2 and D3.3 act like guidance documents and as a frame of reference for future research outputs to be evaluated in the AV preservation space.

Glossary

Term	Definition
CoP	Community of Practice
RO	Research Output

6 References

- [1] Presto4U Consortium, "D3.1: Specification of Assessment Criteria, Metrics, Processes, Datasets and Facilities," 2013.
- [2] Presto4U Consortium, "Description of Work: "European Technology for Digital Media Preservation", " Grant agreement no: 600845, 2011.
- [3] "Blip10000: A social Video Dataset containing SPUG Content for Tagging and Retrieval," 2014.
- [4] "Presto4U Consortium; D3.2: Research Outputs Assessment V1," 2013. [Online].
- [5] van Ossenbruggen, J., Hildebrand, M., and de Boer, V. (2011). Interactive vocabulary alignment. In Proceedings of the 15th International Conference on Theory and Practice of Digital Libraries: Research and Advanced Technology for Digital Libraries , TPD'11, pages 296-307, Berlin, Heidelberg. Springer-Verlag.

Annexes

Please refer to Annex to D3.3: Storage Assessment Results (separate document) for full results of Hardware testing for LTO6 vs Optical Drive.

The Annex presents the results of the testing performed on commercial tools LTO6, Sony Optical Drive and the software used to control this hardware equipment – Front Porch Digital DIVA Software. Due to the commercial nature of these organisations, these test results will be confidential until it is approved by both Sony and Front Porch Digital.

Document information

Delivery Type	Report
Deliverable Number	D3.3
Deliverable Title	Research Outputs Assessments v2
Due Date	31 December 2014
Submission Date	19 December 2014
Work Package	3
Partners	IT Innovation, B&G, INA, RAI, MM, CNR, EURIX, JRS
Author(s)	Bailer Werner, Biscoglio Isabella, Boch Laurent, Borgotallo Roberto, Chakravarthy Ajay, Fabrizio Falchi, Gallo Francesco, Ligos Linda, Laurenson Pip, Melas Panos, Pellegrino Jacopo
Reviewer(s)	Marchetti Eda
Keywords	Assessment, digital preservation, research output
Document Identifier	600845_Deliverable_D3.3_presto4u_18_12_2014(R)
Dissemination level	PU
Document Status	Final
Project Acronym	Presto4U
Project Full Title	European Technology for Digital Audiovisual Media Preservation
Grant Agreement	600845
Project Coordinator	Beeld en Geluid
Contact Details	Sumatrалаan 45, 1217GP Hilversum, The Netherlands. msnyders@beeldengeluid.nl

Document Status Sheet

Version	Delivery Date	Comment	Author
0.1	24/11/2014	First outline created	Ajay Chakravarthy
0.2	26/11/2014	Updated templates and tests for metadata assessment	Werner Bailer
0.3	28/11/2014	Updated templates and tests for technical	Roberto Borgotallo

		metadata extractors	
0.4	10/12/2014	Updated templates and tests for Preservation Systems	Jacopo Pellegrino
0.5	10/12/2014	Updated tests for LTO6 vs Optical Drive	Panos Melas
0.6	14/12/2014	First integrated version for internal review	Ajay Chakravarthy
0.7	16/12/2014	Internal review completed	Eda Marchetti
0.8	16/12/2014	Final version with intro, conclusion, scope, exec summary. Waiting for Amalgame eval	Ajay Chakravarthy
0.9	18/12/2014	Correct minor errors and added missing reference	Ajay Chakravarthy
1.0	19/12/2014	Added Amalgame eval	Werner Bailer