DELIVERABLE

Project Acronym: PREFORMA
Grant Agreement number: 619568
Project Title: PREservation FORMAts for culture information/e-archives

D8.6 Testing Report

Revision: Final 1.20, 27 October 2017

Authors:
Nicola Ferro (UNIPD), Gianmaria Silvello (UNIPD), Erik Buelinckx (KIKIRPA),
Boris Doubrov (VeraPDF), Magnus Geber (Riksarkivet), Klas Jadegljans (Riksarkivet),
Jerôme Martinez (MediaConch), Victor Muñoz (EasyInnova),
Dave Rice (MediaConch), Stefan Rohde-Enslin (SPK), Xavi Tarres (EasyInnova),
Erwin Verbruggen (S&V), Benjamin Yousefi (Riksarkivet), Carl Wilson (VeraPDF)

Reviewers:
Börje Justrell (Riksarkivet), Antonella Fresa (Promoter), Claudio Prandoni (Aedeka)

<table>
<thead>
<tr>
<th>Dissemination Level</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>P Public</td>
<td></td>
</tr>
<tr>
<td>C Confidential, only for members of the consortium and the Commission Services</td>
<td></td>
</tr>
</tbody>
</table>
Revision History

<table>
<thead>
<tr>
<th>Revision</th>
<th>Date</th>
<th>Author</th>
<th>Organisation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>2017-05-05</td>
<td>Nicola Ferro, Gianmaria Silvello</td>
<td>UNIPD</td>
<td>First skeleton</td>
</tr>
<tr>
<td>0.10</td>
<td>2017-05-26</td>
<td>Nicola Ferro, Gianmaria Silvello</td>
<td>UNIPD</td>
<td>Initial draft of the evaluation procedure</td>
</tr>
<tr>
<td>0.20</td>
<td>2017-05-31</td>
<td>Nicola Ferro, Gianmaria Silvello</td>
<td>UNIPD</td>
<td>Initial draft of the ground-truth creation and class refinement</td>
</tr>
<tr>
<td>0.30</td>
<td>2017-06-15</td>
<td>Nicola Ferro, Gianmaria Silvello</td>
<td>UNIPD</td>
<td>Initial draft of the qualitative evaluation</td>
</tr>
<tr>
<td>0.40</td>
<td>2017-07-14</td>
<td>Nicola Ferro, Gianmaria Silvello</td>
<td>UNIPD</td>
<td>Initial skeleton of the quantitative evaluation</td>
</tr>
<tr>
<td>0.50</td>
<td>2017-08-23</td>
<td>Nicola Ferro, Gianmaria Silvello</td>
<td>UNIPD</td>
<td>Refinement of the quantitative evaluation</td>
</tr>
<tr>
<td>0.60</td>
<td>2017-09-08</td>
<td>Nicola Ferro, Gianmaria Silvello</td>
<td>UNIPD</td>
<td>First version circulated to all partners</td>
</tr>
<tr>
<td>0.70</td>
<td>2017-09-11</td>
<td>Nicola Ferro, Gianmaria Silvello</td>
<td>UNIPD</td>
<td>Refinement of the qualitative</td>
</tr>
<tr>
<td>0.80</td>
<td>2017-09-12</td>
<td>Nicola Ferro, Gianmaria Silvello</td>
<td>UNIPD</td>
<td>Second version circulated to all partners</td>
</tr>
<tr>
<td>0.90</td>
<td>2017-09-22</td>
<td>Nicola Ferro, Gianmaria Silvello</td>
<td>UNIPD</td>
<td>Refinement of the quantitative analyses</td>
</tr>
<tr>
<td>1.00</td>
<td>2017-09-25</td>
<td>Nicola Ferro, Gianmaria Silvello</td>
<td>UNIPD</td>
<td>Third version circulated to all partners</td>
</tr>
<tr>
<td>1.10</td>
<td>2017-09-30</td>
<td>Nicola Ferro, Gianmaria Silvello</td>
<td>UNIPD</td>
<td>New cases added to the qualitative analysis</td>
</tr>
<tr>
<td>1.11</td>
<td>2017-10-02</td>
<td>Nicola Ferro, Gianmaria Silvello</td>
<td>UNIPD</td>
<td>Fourth version circulated to all partners</td>
</tr>
<tr>
<td>1.12</td>
<td>2017-10-20</td>
<td>Nicola Ferro, Gianmaria Silvello</td>
<td>UNIPD</td>
<td>Inserted the final evaluation reports of the PREFORMA committee</td>
</tr>
<tr>
<td>1.13</td>
<td>2017-10-21</td>
<td>Nicola Ferro, Gianmaria Silvello</td>
<td>UNIPD</td>
<td>Fifth version circulated to all partners</td>
</tr>
<tr>
<td>1.20</td>
<td>2017-10-27</td>
<td>Nicola Ferro, Gianmaria Silvello</td>
<td>UNIPD</td>
<td>Final version circulated to all partners</td>
</tr>
</tbody>
</table>

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.
## Contents

Executive Summary 7

1 Introduction 9

2 Evaluation Procedure 9
   2.1 Conformance Checking as a Classification Task .......................... 9
   2.2 Measures ................................................................................. 10

3 Ground-truth Creation and Class Refinement 12
   3.1 Creation ................................................................. 12
   3.2 Refinement ............................................................. 14
   3.3 Text Media Type ....................................................... 14
   3.4 Image Media Type .................................................... 17
   3.5 Audio-video Media Type ............................................. 17

4 Quantitative Evaluation 20
   4.1 Text Media Type ....................................................... 20
   4.2 Image Media Type ..................................................... 20
   4.3 Audio-video Media Type ............................................. 20

5 Qualitative Evaluation 27
   5.1 Report on the PREFORMA hand-on seminar in Padua, March 10, 2017 27
   5.2 Report on the PREFORMA hands-on seminars in Hilversum, The Netherlands, January and May 2017 37
   5.3 Report on the PREFORMA hand-on seminar in Barcelona, May 10, 2017 39
   5.4 Report on the PREFORMA hand-on seminar in Stockholm, May 29, 2017 44
   5.5 Report on the PREFORMA hand-on seminar in Quedlinburg, May 29, 2017 46

6 Feedback on the final release of the testing phase and on the EoP report 46
   6.1 Text Media Type ....................................................... 47
      6.1.1 General comments .................................................. 47
      6.1.2 The Conformance Checker ......................................... 47
      6.1.3 Result of compilation of CDP ..................................... 48
      6.1.4 End of Phase Report ............................................... 48
   6.2 Image Media Type ..................................................... 52
      6.2.1 General comments .................................................. 52
      6.2.2 The Conformance Checker ......................................... 53
      6.2.3 Result of compilation of CDP ..................................... 53
      6.2.4 End of Phase Report ............................................... 54
   6.3 Audio/Video Media Type ............................................. 58
      6.3.1 General comments .................................................. 58
6.3.2 The Conformance Checker ....................................... 58
6.3.3 Result of compilation of CDP .................................. 58
6.3.4 End of Phase Report ............................................... 59

A Classes ................................................................. 64
A.1 Text Media Type .................................................... 64
A.2 Image Media Type .................................................. 80
A.3 Audio-video Media Type ........................................... 87

B Created Ground-Truth ............................................... 96
B.1 Text Media Type .................................................... 96
  B.1.1 File List ....................................................... 96
  B.1.2 Ground-truth ................................................. 103
B.2 Image Media Type ................................................ 123
  B.2.1 File List ....................................................... 123
  B.2.2 Ground-truth ................................................. 127
B.3 Audio-video Media Type ......................................... 133
  B.3.1 File List ....................................................... 133
  B.3.2 Ground-truth ................................................. 135

References ............................................................. 137
Executive Summary

Deliverable of D8.6 “Testing Phase” has a twofold goal:

- This deliverable presents the results of the tests conducted on the systems selected in the previous phases of the projects.
- The “Testing Phase” evaluated the tool produced by the suppliers on real experimental collections in order to assess their overall quality for conformance checking.

The document is organized as follows: Section 2 describes the “PREFORMA Evaluation Matrix” tailored for testing the tools selected for the last phase of the project; Section 3 describes the classes of documents that have been removed and refined for the test phase; Section 4 presents the quantitative (classification measures) results of the test phase; Section 5 presents qualitative (focus groups); and, Section 6 reports the comments provided by the members of the PREFORMA Evaluation Committee on the final release of the testing phase (end of August 2017) and on the End of Phase report.
1 Introduction

This release of D8.6 “Testing Phase” has a twofold goal:

- This deliverable presents the results of the tests conducted on the systems selected in the previous phases of the projects.
- The “Testing Phase” evaluated the tool produced by the suppliers on real experimental collections in order to assess their overall quality for conformance checking.

The document is organized as follows: Section 2 describes the “PREFORMA Evaluation Matrix” tailored for testing the tools selected for the last phase of the project; Section 3 describes the classes of documents that have been removed and refined for the test phase; Section 4 presents the quantitative (classification measures) results of the test phase; Section 5 presents qualitative (focus groups); and, Section 6 reports the comments provided by the members of the PREFORMA Evaluation Committee on the final release of the testing phase (end of August 2017) and on the End of Phase report.

2 Evaluation Procedure

2.1 Conformance Checking as a Classification Task

The goal of the PREFORMA conformance checkers is to validate documents against their respective standards. This turns into determining, for each document, whether it is compliant, it suffers from issue 1, issue 2, and so on.

Therefore, we modelled the conformance checking process as a classification task, where you label documents according to their characteristics and each label (compliant, issue 1, issue 2, ...) is a class $C_i$, representing the conformance of or an issue with a document.

In general, classes may intersect, since a document may suffer from multiple issues at the same time, but the compliant class must be a separate one, since you cannot have documents that are compliant and not compliant at the same time, as it is shown in Figure 1.

Figure 1: Conformance checking as a classification task.

In order to evaluate conformance checkers, we relied on the Cranfield paradigm [Cleverdon, 1997], which makes use of experimental collections $C = (D, T, GT)$, where $D$ is a collection of
documents of interest, \( T \) is a set of topics and \( GT \) is the ground-truth which, for each document \( d \in D \) and topic \( t \in D \), determines the relevance of document \( d \) to topic \( t \). In the classification context, this paradigm is instantiated considering the classes \( C_i \) as topics and the ground-truth is given by the correct labels assigned to each document \( d \) [Sebastiani, 2002].

In terms of the approach proposed by [Duretec et al., 2015], we have that: the motivating comparison is given by the need of assessing conformance checkers; the task sample is defined by the identified classes \( C_i \); the gathered documents, and the ground-truth; the performance measures are described in Section 2.2.

A more detailed description of this framework is available in [Ferro, 2016; Ferro et al., 2016].

### 2.2 Measures

Considering that we frame conformance checking as a classification task, it becomes natural to evaluate it according to the confusion matrix [Sokolova and Lapalme, 2009] shown in Figure 2.

Recall from Section 2.1 and Figure 1 that each class \( C_i \) represents a possible mis-conformance with respect to a reference standard with the exception of the class \( C_0 \) which represents documents fully conforming to the standard.

In the confusion matrix:

- **True Positive (TP):** it is the set of documents that a conformance checker has correctly labeled as belonging to class \( C_i \);

- **True Negative (TN):** it is the set of documents that a conformance checker has correctly labeled as not belonging to class \( C_i \);

- **False Positive (FP):** it is the set of documents that a conformance checker has incorrectly labeled as belonging to class \( C_i \);

- **False Negative (FN):** it is the set of documents that a conformance checker has incorrectly labeled as not belonging to class \( C_i \).

<table>
<thead>
<tr>
<th>Class ( C_i )</th>
<th>Ground-Truth</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td>Conformance Checker</td>
<td>Positive</td>
<td>True Positive (TP)</td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>False Negative (FN)</td>
</tr>
</tbody>
</table>

Figure 2: Confusion matrix for the evaluation of conformance checkers for each class \( C_i \).
Note that what we mean by the confusion matrix of Figure 2 changes if we are considering $C_0$, i.e. the class representing a compliant document, or a generic $C_i$, $i \neq 0$, i.e. a class representing an issue within a document.

In the case of $C_0$, $TP_0$ is the set of compliant documents correctly identified as compliant; $TN_0$ is the set of not compliant documents correctly identified as not compliant; $FP_0$ is the set of not compliant documents incorrectly identified as compliant; and, $FN_0$ is the set of compliant documents incorrectly identified as not compliant.

In the case of $C_i$, $i \neq 0$, $TP_i$ is the set of not compliant documents because of issue $i$ correctly identified as suffering from issue $i$; $TN_i$ is the set of documents correctly identified as not suffering from issue $i$; $FP_i$ is the set of documents incorrectly identified as suffering from issue $i$; $FN_i$ is the set of not compliant documents because of issue $i$ but incorrectly identified as not suffering from issue $i$.

Note that the impact of FP and FN is different in the case we are considering $C_0$ or a generic $C_i$, $i \neq 0$. In the case of $C_0$, FPs are the worst error for a conformance checker, since they are not conforming documents marked as compliant and thus allowed to proceed in the preservation chain, possibly causing issues in the long term; on the other hand, FNs are a less sever error, since they are compliant documents marked as not compliant which will require some additional work for further checks and fixes (actually not necessary) but, eventually, they will have a chance to go ahead in the preservation chain. In the case of $C_i$, $i \neq 0$, FNs are the worst error for a conformance checker, since they are undetected not compliant documents thus allowed to proceed in the preservation chain, possibly causing issues in the long term; on the other hand FPs are just a kind of “false alarm”, which will require some additional work for further checks and fixes (actually not necessary) but, eventually, they will have a chance to go ahead in the preservation chain.

Therefore, we will rely on evaluation measures able both to give a general account of conformance checkers performances and to deal with this duality between FNs and FPs:

- **accuracy**: measures the overall effectiveness [Sokolova and Lapalme, 2009] of a conformance checker as

\[
\text{Accuracy}_i = \frac{|TP_i| + |TN_i|}{|TP_i| + |TN_i| + |FP_i| + |FN_i|}
\]

- **area under the curve (AUC)**: measures the ability of a conformance checker to avoid false classification [Fawcett, 2006; Sokolova and Lapalme, 2009] as

\[
\text{AUC}_i = \frac{1}{2} \left( \frac{|TP_i|}{|TP_i| + |FN_i|} + \frac{|TN_i|}{|TN_i| + |FP_i|} \right)
\]

In order to obtain a single score for each conformance checker across all the categories $C_i$, we will use a macro-averaging approach [Sebastiani, 2002], which computes the arithmetic mean of the above measures over all the categories $C_i$.

Moreover, as explained in Section 2.1, since a document cannot be compliant and not compliant at the same time, the class $C_0$ of the compliant documents must be separate from any other class $C_i$ representing a possible issue of a document, i.e. $C_0 \cap C_i = \emptyset \forall i, i \neq 0$. As a consequence, assuming perfect classification, i.e. no FP or FN happen, it should be $TP_0 \cap TP_i = \emptyset \forall i, i \neq 0$, i.e. there must
be no intersection between the TP documents attributed to $C_0$ and those attributed to other classes $C_i$. Since classification is typically not perfect, it should hold that $(TP_0 \cup FP_0) \cap (TP_i \cup FP_i) = \emptyset \forall i, i \neq 0$, i.e. the documents that a conformance checker correctly or incorrectly attributes to $C_0$ should have no intersection with the documents it correctly or incorrectly attributes to other classes $C_i$. Another consequence is that $TN_0 \cup FN_0 = \bigcup_{i=1}^{N} (TP_i \cup FP_i)$, i.e. the documents correctly or incorrectly marked as not compliant by a conformance checkers must have been attributed to some other class $C_i$ by the same conformance checker.

Therefore, we can introduce an additional overall performance measure, called consistency, which assesses the ability of a conformance checker to adhere to the above constraint of separation of $C_0$ from the other classes:

$$\text{Consistency}_i = 1 - \frac{|(TP_0 \cup FP_0) \cap (TP_i \cup FP_i)|}{|(TP_i \cup FP_i)|} = 1 - \frac{|C_0 \cap C_i|}{|C_i|}$$

Note that consistency is different from the evaluation measures typically used in classification [Ferri et al., 2009; Sebastiani, 2002; Sokolova and Lapalme, 2009] or clustering [Amigó et al., 2009, 2013] and serves the specific purpose of assessing the degree of separation between the compliant and not-compliant classes.

3 Ground-truth Creation and Class Refinement

3.1 Creation


For each class it is specified:

- a unique identifier;
- a short name;
- a brief description.

For each media type a domain expert group has been established and was in charge of defining the list of classes. Each domain expert group is constituted as follows:

- 1 evaluation expert, i.e. an expert of organization of evaluation activities according to the Cranfield paradigm who oversees the classes definition process and facilitates the discussion within the group;

\(^1\)http://www.ffmpeg.org/~michael/ffv1.html
\(^2\)http://www.matroska.org/
• 2 experts from memory institutions, i.e. one technical and one domain expert representing the viewpoint of the memory institutions which are the main stakeholders of the project;

• 1 expert from suppliers, i.e. one technical expert representing the viewpoint of the suppliers which are the other main stakeholders of the project.

The composition of the domain expert group ensure a fair representation of all the different viewpoints involved in the PREFORMA project. The domain expert groups, for each media type, are as follow:

• text media type:
  – Erik Buelinckx
  – Boris Doubrov
  – Magnus Geber
  – Eva McEneaney
  – Marju Niinemaa
  – Benjamin Yousefi
  – Carl Wilson

• image media type:
  – Erik Buelinckx
  – Peter Fornaro
  – David Iglesias
  – Klas Jadeglans
  – Uwe Kühhirt
  – Bert Lemmens
  – Víctor Munoz
  – Bengt Neiss
  – Peter Pharow
  – Stefan Rohde-Enslin
  – Xavi Tarres
  – Christian Weigel

• audio-video media type:
  – Anna Kasimati
  – Uwe Kühhirt
  – Bert Lemmens
Appendix A reports the class which has been actually used in the evaluation and which has been updated with respect to the original list reported in deliverable 8.1R2 [Ferro et al., 2016], as explained in the next section.

The groups of experts have also been responsible for the actual creation of the ground-truth for each class, i.e. they determined a set of files belonging to each class. Appendix B reports the ground-truth created for each media type and which has been used to evaluate the systems.

3.2 Refinement

During the creation of the ground-truth and the initial iterations of the evaluation phase, the expert groups and the PREFORMA suppliers have carefully revised the list of classes initially defined in deliverable 8.1R2 [Ferro et al., 2016] and they determined a few classes that had to be eliminated.

The following sections list the removed classes for each media type along with the motivation for their removal; note that in the case of the audio-video media type changes in the draft standards are reflected in the updated class list.

3.3 Text Media Type

- Class TC001 – Annotation FileAttachment
  - description: Has File attached to the PDF/A document.
  - PDF/A version: 1
  - conformance level: ba
  - class type: Conformance
  - severity: 5
  - motivation for removal: (almost) complete overlap with TC002.

- Class TC011 – Image Alternative/Proxy
  - PDF/A version: 1-3
  - conformance level: bua
  - class type: Conformance
– severity: 5
– motivation for removal: The class cannot be reproduced because of missing tool set.

• Class TC015 – Import/Link to External Resource
  – description: Has links to an external resource (rather than embedding it) such as external File specifications and reference XObjects.
  – PDF/A version: 1-3
  – conformance level: bua
  – class type: Conformance
  – severity: 5
  – motivation for removal: The class cannot be reproduced because of missing tool set.

• Class TC057 – Embed fonts Type 3
  – description: Does not have Font Type 3 embedded. A Type 3 font dictionary defines the font; font dictionaries for other fonts simply contain information about the font and refer to a separate font program for the actual glyph descriptions (PDF reference, sec. 9.6.5). MAY NOT BE NECESSARY. DOCUMENT: WHY NOT FEASIBLE OR PRACTICAL OR NOT POSSIBLE. - Font Type 3 as defined to function will always be embedded when used.
  – PDF/A version: 1-3
  – conformance level: bua
  – class type: Conformance
  – severity: 5
  – motivation for removal: The class cannot be tested because Font Type 3 as defined: embedding is a function of use. The class was supposed to be dropped and has accidently been included

• Class TC061 – Character encoding symbolic TrueType fonts
  – description: Has an Encoding entry in the font dictionary for any Symbolic TrueType fonts, or the ‘cmap’ table in the embedded font program neither contains exactly one encoding nor contains at least the Microsoft Symbol (3,0 ’ Platform ID=3, Encoding ID=0) encoding.
  – PDF/A version: 2-3
  – conformance level: bua
  – class type: Conformance
  – severity: 5
  – motivation for removal: The class cannot be reproduced because of unclear definition and missing instructions.
• Class TC063 – **Conformance Level B**
  - *description*: Does not have functionality required for Conformance Level B.
  - *PDF/A version*: 1-3
  - *conformance level*: b
  - *class type*: Conformance
  - *severity*: 5
  - *motivation for removal*: This class is too generic and not correctly specified.

• Class TC072 – **Attached document is XML.**
  - *description*: Does not have XML file attached.
  - *PDF/A version*: 3
  - *conformance level*: Any
  - *class type*: Policy
  - *severity*: 5
  - *motivation for removal*: This class is too generic and not correctly specified.

• Class TC076 – **Presence of Javascript**
  - *description*: Does not have javascript
  - *PDF/A version*: Any
  - *conformance level*: Any
  - *class type*: Policy
  - *severity*: 5
  - *motivation for removal*: This class is too generic and not correctly specified.

• Class TC087 – **Digital signature should be present**
  - *description*: The document is not digitally signed
  - *PDF/A version*: Any
  - *conformance level*: Any
  - *class type*: Policy
  - *severity*: 5
  - *motivation for removal*: This class is too generic and not correctly specified.

• Class TC088 – **Outline / bookmarks present**
  - *description*: The document does not contain an outline
  - *PDF/A version*: Any
– conformance level: Any
– class type: Policy
– severity: 5
– motivation for removal: This class is too generic and not correctly specified.

3.4 Image Media Type

• Class IC044 – Size of the uncompressed TIFF-file
  – description: Size in the range [X, Y] Mb (depending on the size of the analogue object)
  – class type: Policy
  – severity: 3
  – motivation for removal: since the range X - Y is not fixed, as it depends on memory institutions conditions, it makes no sense to create generic TIFFs related to this class.

3.5 Audio-video Media Type

• Class AVC009 – Header Elements in Element ID length range
  – description: Element ID (descending from Root Element) lengths must be less than or equal to 4. [EBML/HEADER-ELEMENTS-WITHIN-MAXIDLENGTH]
  – class type: Conformance
  – severity: 1
  – motivation for removal: Deprecated. It would require an invalid EBML Schema to create such a file.

• Class AVC010 – Elements in Element ID length range
  – description: Element ID (descending from Root Element) lengths must be less than or equal to EBMLMaxIDLength. [EBML/ELEMENTS-WITHIN-MAXIDLENGTH]
  – class type: Conformance
  – severity: 1
  – motivation for removal: Deprecated. After recent specification clarifications, it would require an invalid EBML Schema in order to support this error.

• Class AVC013 – EBML vint efficiency
  – description: Section 2.2 IDs are always encoded in their shortest form e.g. 1 is always encoded as 0x81 and never as 0x4001." The bits following the Element ID’s Length Descriptor are not more than (8 - $bit-length-of-length-descriptor) successive 0 bits i.e. vint is expressed as efficiently as feasible." [EBML/EBML-VINT-EFF]
  – class type: Conformance
- **severity**: 1
- **motivation for removal**: Deprecated. It would require an invalid Matroska definition to support.

- **Class AV C014 – Element ID Registered**
  - **description**: Ensure MKV Element ID is registered in specdata.xml (as of Dec. 13 2014 this is 224 registered Element IDs) [EBML/MKV-KNOWN-ELEM]
  - **class type**: Conformance
  - **severity**: 1
  - **motivation for removal**: Deprecated by NO-JUNK-IN-FIXEDSIZE_MATROSKA and EBML-ELEMENT-VALID-PARENT. This test would be redundant, any known Element would have a known Parent Element and another unrecognized Element would be considered junk which is only conditionally valid is elements that allow unknown sizes.

- **Class AV C016 – Element Size Byte Length Limit**
  - **description**: Section 2.3: The EBML element data size is encoded as a variable size integer with by default widths up to 8.” The first eight bits of any Element Size may not start with 0b00000000.” [EBML/EBML-ELEM-SIZE-CAP]
  - **class type**: Conformance
  - **severity**: 1
  - **motivation for removal**: Deprecated by ELEMENTS-WITHIN-MAXSIZELENGTH.

- **Class AV C017 – Element Size Unknown**
  - **description**: only Master Elements may be unknown size [EBML/EBML-ELEM-SIZE-UNK]
  - **class type**: Conformance
  - **severity**: 1
  - **motivation for removal**: Deprecated by EBML-ELEM-UNKNOWN-SIZE.

- **Class AV C018 – Element Data within Size Limits**
  - **description**: test EBML Schema size restrictions per element [EBML/EBML-WITHIN-SIZE-LIMIT]
  - **class type**: Conformance
  - **severity**: 1
  - **motivation for removal**: Deprecated by EBML-ELEMENT-IN-SIZE-RANGE.

- **Class AV C023 – EBML CRC Element must use a valid length**
– **description**: CRC values are required to be 4 bytes in length. [EBML/EBML-CRC-LENGTH]
  – **class type**: Conformance
  – **severity**: 1
  – **motivation for removal**: Deprecated by EBML-ELEMENT-IN-SIZE-RANGE.

• **Class AVC026 – EBML Elements used correlate to DocTypeReadVersion**
  – **description**: The EBMLDocTypeReadVersion should not be lower than the minver of Elements essential to proper playback. [EBML/EBML-DOCTYPEREADVERSION-COHERANT]
  – **class type**: Conformance
  – **severity**: 1
  – **motivation for removal**: The setting of DocTypeReadVersion is partly to the judgement of the author of the file in regards to what features of the file are important.

• **Class AVC027 – Matroska Segment Element must use a valid length**
  – **description**: The length of the value of Segment UID must be 16 bytes. [Matroska/MKV-SEGMENT-UID-LENGTH]
  – **class type**: Conformance
  – **severity**: 1
  – **motivation for removal**: Deprecated by EBML-ELEMENT-IN-SIZE-RANGE.

• **Class AVC031 – Matroska Boolean Elements are valid**
  – **description**: Some elements are defined as boolean but expressed in unsigned integer; verify that they are valid. [Matroska/MKV-VALID-BOOLEANS]
  – **class type**: Conformance
  – **severity**: 1
  – **motivation for removal**: Deprecated by EBML-ELEMENT-VALID-RANGE.

• **Class AVC046 – initial_state_delta**
  – **description**: initial_state_deltas incoherency [FFV1/FFV1-HEADER-initial_state_delta]
  – **class type**: Conformance
  – **severity**: 1
  – **motivation for removal**: All data is valid initial_state_delta, as the specification includes a mask.

• **Class AVC060 – PCM is valid**
  – **description**: some data is there [FFV1/PCM-IS-VALID]
Table 1: Average values of accuracy, AUC and consistency for the text media type. The minimum value is 0 and the maximum is 1.

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>AUC</th>
<th>Consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9812</td>
<td>0.8495</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

- **class type**: Conformance
- **severity**: 1
- **motivation for removal**: All data is valid PCM

### 4 Quantitative Evaluation

#### 4.1 Text Media Type

In Table 1 we can see the values of accuracy, AUC and consistency averaged over all the text classes. The results are pretty good since accuracy is very close to the maximum value, whereas AUC could be further improved. Consistency is 95%, hence there are some documents which are complaint and not complaint at the same time.

In Table 2 we can see the detailed results achieved for each test class. This table points us to the few classes where there are some mis-classification, we provide also Figure 3 for the accuracy and Figure 4 for the AUC to better grasp the performance differences between the classes.

**VeraPDF and Text Experts:** Add possible explanations for the low performing classes.

#### 4.2 Image Media Type

In Table 3 we can see the values of accuracy, AUC and consistency averaged over all the image classes.

In Table 4 we can see the detailed results achieved for each test class. This table gives us a more accurate view over the accuracy, AUC and consistency of all the classes at hand where we can see that the checker behaves perfectly; we also provide Figure 5 for the accuracy and Figure 6 for the AUC.

#### 4.3 Audio-video Media Type

In Table 5 we can see the values of accuracy, AUC and consistency averaged over all the audio-video classes. The results are pretty good since accuracy and AUC are very close to one that would mean perfect classification for all classes. Consistency is one, thus there is no document which is complaint and not complaint at the same time.

In Table 6 we can see the detailed results achieved for each test class. This table points us to the few classes where there are some mis-classification, the most critical class, is AVC061 where the conformance checker get 96% of accuracy; this class asks to verify if a file is at least version 4 of the Matroska standard. As noted above, the Matroska standard has been evolving during the project.
Table 2: Class-by-class values of accuracy, AUC and consistency for the text media type. The minimum value is 0 and the maximum is 1.

<table>
<thead>
<tr>
<th>Class</th>
<th>Accuracy</th>
<th>AUC</th>
<th>Consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td>TC000</td>
<td>0.8660</td>
<td>0.6514</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC002</td>
<td>0.9823</td>
<td>0.4970</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC003</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC004</td>
<td>0.9790</td>
<td>0.9894</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC005</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC006</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC007</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC008</td>
<td>0.9881</td>
<td>0.7470</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC009</td>
<td>0.9880</td>
<td>0.9939</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC010</td>
<td>0.9910</td>
<td>0.4985</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC012</td>
<td>0.9760</td>
<td>0.9878</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC013</td>
<td>0.9910</td>
<td>0.9955</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC014</td>
<td>0.9521</td>
<td>0.8772</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC016</td>
<td>0.9341</td>
<td>0.8432</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC017</td>
<td>0.9734</td>
<td>0.6910</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC019</td>
<td>0.9552</td>
<td>0.7304</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC021</td>
<td>0.9880</td>
<td>0.9939</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC022</td>
<td>0.9911</td>
<td>0.6250</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC024</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC025</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC026</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC027</td>
<td>0.9760</td>
<td>0.9879</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC030</td>
<td>0.9970</td>
<td>0.9985</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC033</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC035</td>
<td>0.9790</td>
<td>0.9894</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC037</td>
<td>0.9822</td>
<td>0.4985</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC038</td>
<td>0.9852</td>
<td>0.5833</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC039</td>
<td>0.9881</td>
<td>0.6985</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC040</td>
<td>0.9580</td>
<td>0.9786</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC042</td>
<td>0.9911</td>
<td>0.9681</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC043</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC044</td>
<td>0.9820</td>
<td>0.8674</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC045</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC047</td>
<td>0.9823</td>
<td>0.9454</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC048</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC049</td>
<td>0.9700</td>
<td>0.9848</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC050</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC051</td>
<td>0.9880</td>
<td>0.8288</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC052</td>
<td>0.9970</td>
<td>0.9985</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC054</td>
<td>0.9852</td>
<td>0.4985</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC055</td>
<td>0.9670</td>
<td>0.9832</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC058</td>
<td>0.9850</td>
<td>0.9925</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC059</td>
<td>0.9910</td>
<td>0.9897</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC060</td>
<td>0.9640</td>
<td>0.9818</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC062</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC064</td>
<td>0.9410</td>
<td>0.4747</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC065</td>
<td>0.9910</td>
<td>0.9904</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC070</td>
<td>0.9822</td>
<td>0.4955</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC071</td>
<td>0.9735</td>
<td>0.4955</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC075</td>
<td>0.9554</td>
<td>0.4820</td>
<td>1.0000</td>
</tr>
<tr>
<td>TC077</td>
<td>0.9730</td>
<td>0.4896</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
Figure 3: Stem plot of the class accuracy for the text media type. The red horizontal line indicates the average.

Table 3: Average values of accuracy, AUC and consistency for the image media type. The minimum value is 0 and the maximum is 1.

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>AUC</th>
<th>Consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0000</td>
<td>1.000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
Figure 4: Stem plot of the class AUC for the text media type. The red horizontal line indicates the average.
Table 4: Class-by-class values of accuracy, AUC and consistency for the image media type. The minimum value is 0 and the maximum is 1.

<table>
<thead>
<tr>
<th>Class</th>
<th>Accuracy</th>
<th>AUC</th>
<th>Consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td>IC000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC001</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC002</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC003</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC004</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC005</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC006</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC007</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC008</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC009</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC010</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC011</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC012</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC013</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC014</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC015</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC016</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC017</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC018</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC019</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC020</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC021</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC022</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC023</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC024</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC025</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC026</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC027</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC028</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC029</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC030</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC031</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC032</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC033</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC034</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC035</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC036</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC037</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC038</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC039</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC040</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC041</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC042</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>IC043</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
Figure 5: Stem plot of the class accuracy for the image media type. The red horizontal line indicates the average.

Table 5: Average values of accuracy, AUC and consistency for the audio-video media type. The minimum value is 0 and the maximum is 1.

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>AUC</th>
<th>Consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9977</td>
<td>0.9621</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
Figure 6: Stem plot of the class AUC for the image media type. The red horizontal line indicates the average.
and its changes could have affected the performances of the checker in some cases. All the other classes show perfect classification or an almost negligible error.

Audio-Video Experts: Add further possible explanations for the low performing classes.

5 Qualitative Evaluation

Since when the PREFORMA prototypes have been completed and stabilized, the PREFORMA project organized a series of hands-on sessions and training seminars to explain to the participants what does conformance checking mean, why is file format validation so important in long-term digital preservation, how to create their own policy profiles and how to download, install, configure and use the conformance checker to analyze their files.

These workshops/seminars invited archivists/conservators/librarians to bring their files and analyze them with the PREFORMA tools. At the end of the workshop, they understood which are the main issues related to digital preservation and file formats validation at many memory institutions, check whether their files conform to the specifications of the standards, and learnt how to create a policy profile that allows them to check if their files are compliant with the acceptance criteria for their digital repository.

5.1 Report on the PREFORMA hand-on seminar in Padua, March 10, 2017

The first of PREFORMA hands-on sessions has been organized on 10 March 2017 in Padua (Italy) in combination with the Innovation Workshop. The event was very successful and brought together more that 20 librarians with technical knowledge and IT staff dealing with long-term preservation from the universities of Padua, Venice Ca’ Foscari and IUAV (Venice).

The programme of the session was the following:

- 9.00 Introduction (Bert Lemmens, Claudio Prandoni)
  - Tour de table asking the participants where they come from and the level of digital preservation at their institutions.
  - Short overview of the importance of conformance checking in digital preservation.
  - Brief introduction to the PREFORMA tools, showing what they can do and how to use them.

- 9.30 Exercise 1. Installing and running the checkers
  - Install the tool
  - Run the checker
  - Test your files
  - Use a predefined policy
  - Make a report

- 10.15 Exercise 2. Reading the report
Table 6: Class-by-class values of accuracy, AUC and consistency for the audio-video media type. The minimum value is 0 and the maximum is 1.

<table>
<thead>
<tr>
<th>Class</th>
<th>Accuracy</th>
<th>AUC</th>
<th>Consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVC000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC001</td>
<td>0.9863</td>
<td>0.7500</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC002</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC003</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC004</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC005</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC006</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC007</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC008</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC009</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC010</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC011</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC012</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC013</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC014</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC015</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC016</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC017</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC018</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC019</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC020</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC021</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC022</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC023</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC024</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC025</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC026</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC027</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC028</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC029</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC030</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC031</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC032</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC033</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC034</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC035</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC036</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC037</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC038</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC039</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC040</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC041</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC042</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC043</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC044</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC045</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC046</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC047</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC048</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC049</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC050</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC051</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC052</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC053</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC054</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC055</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC056</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC057</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC058</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC059</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC061</td>
<td>0.9600</td>
<td>0.4932</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC062</td>
<td>0.9863</td>
<td>0.7500</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC063</td>
<td>0.9863</td>
<td>0.7500</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC064</td>
<td>0.9863</td>
<td>0.7500</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC065</td>
<td>0.9863</td>
<td>0.7500</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC066</td>
<td>0.9863</td>
<td>0.7500</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC067</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC068</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AVC069</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
Figure 7: Stem plot of the class accuracy for the audio/video media type. The red horizontal line indicates the average.
Figure 8: Stem plot of the class AUC for the audio/video media type. The red horizontal line indicates the average.
– What are the reports available?
– What is in the report?

- 10.45 Break
- 11.00 Introduction to file policies (Borje Justrell)
  - Introduction: What are file policies? What do you need them for? (Borje)
  - What properties should be retained in your file?
  - Create a policy profile for your organisation
  - How to communicate this to your organisation (management, digitisation services, registrars)
- 11.15 Exercise 3. Defining your policy
- 12.00 Survey & feedback
  - Feedback discussion
  - Fill in an evaluation questionnaire

In Figure 9 and Figure 10 we show two pictures of the hands-on session held in Padua. In the first one, we see the introduction to the session by Claudio Prandoni and in the second the actual hands-on session where the participants are carrying out the assigned tasks.

Figure 9: Picture of the Hands-on session in Padua. The introduction by Claudio Prandoni.
At the end of the session, fifteen questionnaires with general and specific questions about PREFORMA and the conformance checkers were collected. Eleven participants were interested in the text format, twelve in the images and nine in the audio/video format;

The participants were asked:

- “Is there any other file format for which you would be interested in having a conformance checker, besides those already covered in PREFORMA (PDF/A, TIFF, Matroska/FFv1)?” All answered “No” and one wanted also the TXT format to be covered.

- How would you envisage to use the results of the open source projects in your legacy environment and/or in your digital archiving and preservation initiatives? Eight answered “Integration in legacy system” and six “Standalone service via web”.

- Do you expect that the PREFORMA’s results will have a positive impact on the workflow of your institution/organisation? Five answered “yes” because “It could improve the workflow, and relations with other institutions and agencies”, “To test the tools and to find a possible integration with our system”, “to
increase the awareness of the issue” and “to increase discussion about policies and standard-
isation”; all the others answered “too early to say”.

All the participants tested all the conformance checker and we asked how they would rate their
experience in using the tool(s) in terms of performance, usability, potential and other related aspects.
The opinion were expressed with a number from 1 (very poor) to 5 (very good) and with a free text.
The answers are reported in Table 7 and visualized with an histogram in Figure 11.

Another question was: “How would you rate your experience in using the tool(s) in terms of
specific functionality?” The answers are reported in Table 8 and visualized with an histogram in
Figure 12.

To the question “What’s the one single thing that you learnt today?”, the participants answered
(one did not answer):

- The importance of some details, too often forgotten.
- Met PREFORMA knowledge.
- Creating policies.
- How to create policy for AV.
- There is a good and interesting tool which deserves to be analysed and tested deeper.
- You can implement your own checking rules.
- That a TIFF is not a perfect file just because its extension is .tif
- There is a community working to answer my doubts.
- Application of correct parameters for the long-term preservation.
- More consciousness of the obsolescence.
Table 7: How would you rate your experience in using the tool(s) in terms of performance, usability, potential and other related aspects? Please, express your opinion with a number from 1 (very poor) to 5 (very good). [Padua hands-on session]

<table>
<thead>
<tr>
<th>Innovation potential</th>
<th>Performance / Precision</th>
<th>Usability</th>
<th>Support / Documentation</th>
<th>Scalability / Extensibility</th>
<th>Interoperability</th>
<th>Licenses</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>2.5</td>
<td>3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td>2.5</td>
<td>3</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2.5</td>
<td>3</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>4</td>
<td>-</td>
</tr>
</tbody>
</table>
Figure 11: How would you rate your experience in using the tool(s) in terms of performance, usability, potential and other related aspects? Please, express your opinion with a number from 1 (very poor) to 5 (very good). [Padua hands-on session]
Figure 12: How would you rate your experience in using the tool(s) in terms of performance, usability, potential and other related aspects? Please, express your opinion with a number from 1 (very poor) to 5 (very good). [Padua hands-on session]
5.2 Report on the PREFORMA hands-on seminars in Hilversum, The Netherlands, January and May 2017

At the Netherlands Institute for Sound and Vision, MediaConch was trial-ed and shown on three separate occasions. The first time was in January 2017 during the Winter School for Audiovisual Archiving, in which trainer Peter Bubestinger showed the capabilities of MediaConch to check a file’s contents. Participants were encouraged to walk through the elements of a format policy by means of the software’s policy creation interface.

![Figure 13: MediaConch in use during the Winter School for Audiovisual Archiving 2017. Photo credit: Sebastiaan ter Burg, CC BY.](image)

The second occasion was in May 2017, when at the AVA_Net Symposium (http://www.beeldengeluid.nl/avanet-symposium) Bert Lemmens, Emanuel Lorrain, and Erwin Verbruggen conducted an intensive workshop, outlining the use case for MediaConch and testing the software with 8 participants. Participants ranged from digitisation providers to audiovisual, technical staff and regional archivists. For most of the participants, it was a first-time encounter with the software. Technical staff from Sound and Vision commented on the usefulness of having a policy checker to certify materials in the pre-ingest phase - as they thus far had no affordable tools to perform this QC step before ingest into the collection.

The third occasion was an expert meeting set up in the framework of the The Reel Thing XL symposium in Amsterdam, also in May 2017. The meeting gathered experts from the film archiving domain across Europe.
In order to give the film (and wider audio-visual) archiving community a firmer grasp on the potential of open source conformance tools for preservation purposes, the session was organised as a mix of presentations, demonstrations, hands-on exercises and an open discussion with those present, with the aim to:

1. introduce the current status of standardization work on open formats MKV (wrapper) and FFv1 (codec) in the context of the PREFORMA project;
2. share use cases from archives who have implemented these formats in their preservation activities;
3. share knowledge on using the tools to get a firm grasp on these formats’ usage within a film & av preservation environment;
4. share thoughts and opinions amongst the professionals present on their needs & use cases for further development of said standards & tools.

Erwin Verbruggen (Sound and Vision) introduced the FFV1/MKV development and CELLAR update, after which Eva Verdooldt, Noortje Verbeke, and Matthias Priem (VIAA) talked about the process of selecting an archive master for their mixed AV collection. Jérôme Martinez (MediaArea) did a show and tell of the MediaConch toolset, followed by Kieran O’SLeary (IFI) who demonstrated how the Irish Filmstitute adopted FFV1 for DPX files. Reto Kromer (reto.ch) discussed getting to grips with FFmpeg and Bash in a digital preservation workflow. Stephen McConnachie (BFI) guided an open discussion with the 30 participants about the use of these open source in the film archiving domain.

Figure 14: Room overview during the The Reel Thing XL Expert Session. Photo credit: Erwin Verbruggen, CC BY.
5.3 Report on the PREFORMA hand-on seminar in Barcelona, May 10, 2017

This seminar took place in the premises of COBDC on the 10th of May and the session was conducted by Sònia Oliveras from the Girona City Council and Xavier Tarrés and Víctor Muñoz from Easy Innova. The attendees, who have large amounts of TIFF files, came mostly from local and national memory institutions and weren’t aware of any file conformance checker. The tools offered by PREFORMA project were the first solution in order to solve the file format conformance that they know.

It was the opportunity to make a hands on session with members of the Official Association of Librarians (Collegi Oficial de Bibliotecaris i Documentalistes de Catalunya - COBDC) and with the DPF Manager software in order to check TIFF files.

Five questionnaires with general and specific questions about PREFORMA and the conformance checkers were collected at the end of the seminar. One participant came from a public library, one from public administration, one was an electronic administration management consultant and two worked in a cultural heritage institution. Three of the participants were interested in all the three media types covered by PREFORMA, whereas one only in text and images and another only in images and audio/video.

The participants were asked:

- “Is there any other file format for which you would be interested in having a conformance checker, besides those already covered in PREFORMA (PDF/A, TIFF, Matroska/FFv1)?”
  Four answered “No” and one wanted also the JPEG format to be covered.

- How would you envisage to use the results of the open source projects in your legacy environment and/or in your digital archiving and preservation initiatives?
  Three answered “Web” and two (from the cultural heritage domain) “Integration to legacy tools”.

- Do you expect that the PREFORMA’s results will have a positive impact on the workflow of your institution/organisation?
  Two answered “yes”, in particular to check TIFF files; one answered “no” (electronic adm. consultant), one did not answer and one said it is too early to know.

All the participants tested all the conformance checker and we asked how they would rate their experience in using the tool(s) in terms of performance, usability, potential and other related aspects. The opinion were expressed with a number from 1 (very poor) to 5 (very good) and with a free text. The answers are reported in Table 9 and visualized with an histogram in Figure 15.

Another question was: “How would you rate your experience in using the tool(s) in terms of specific functionality?” The answers are reported in Table 10 and visualized with an histogram in Figure 16.

To the question “What’s the one single thing that you learnt today?”, the participants answered (one did not answer):

- The digital preservation could be possible because we knew a right tools.
Table 8: How would you rate your experience in using the tool(s) in terms of specific functionality? Please, express your opinion with a number from 1 (very poor) to 5 (very good). [Barcelona hands-on session]

<table>
<thead>
<tr>
<th>Download, installation and configuration</th>
<th>User Interface</th>
<th>Command Line Interface</th>
<th>Conformance checking</th>
<th>Creation and checking of a policy</th>
<th>Interpretation of the results</th>
<th>Metadata fixing</th>
<th>Documentation</th>
<th>Web-based version</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>5</td>
<td>-</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>-</td>
<td>3</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>-</td>
<td>5</td>
<td>5</td>
<td>-</td>
<td>5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3.5</td>
<td>4.5</td>
<td>-</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>-</td>
<td>5</td>
<td>4</td>
<td>-</td>
<td>4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>-</td>
<td>5</td>
<td>4</td>
<td>-</td>
<td>4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3.5</td>
<td>4.5</td>
<td>-</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>-</td>
<td>5</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3.5</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>-</td>
<td>4</td>
<td>3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 9: How would you rate your experience in using the tool(s) in terms of performance, usability, potential and other related aspects? Please, express your opinion with a number from 1 (very poor) to 5 (very good). [Barcelona hands-on session]

<table>
<thead>
<tr>
<th>Innovation potential</th>
<th>Performance / Precision</th>
<th>Usability</th>
<th>Support / Documentation</th>
<th>Scalability / Extensibility</th>
<th>Interoperability</th>
<th>Licenses</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>5</td>
<td>5</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Figure 15: How would you rate your experience in using the tool(s) in terms of performance, usability, potential and other related aspects? Please, express your opinion with a number from 1 (very poor) to 5 (very good). [Barcelona hands-on session]

Table 10: How would you rate your experience in using the tool(s) in terms of specific functionality? Please, express your opinion with a number from 1 (very poor) to 5 (very good). [Barcelona hands-on session]

<table>
<thead>
<tr>
<th>Download, installation and configuration</th>
<th>User Interface</th>
<th>Command Line Interface</th>
<th>Conformance checking</th>
<th>Creation and checking of a policy</th>
<th>Interpretation of the results</th>
<th>Metadata fixing</th>
<th>Documentation</th>
<th>Web-based version</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>4</td>
<td>-</td>
<td>5</td>
<td>5</td>
<td>-</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

PREFORMA Deliverable 8.6 page [41] of [138]
Figure 16: How would you rate your experience in using the tool(s) in terms of specific functionality? Please, express your opinion with a number from 1 (very poor) to 5 (very good). [Barcelona hands-on session]
• Identifying and fix file formats and know a new software to do that.
• There are other formats for digital preservation, apart of PDF/A. There is a software that can help to the adminstrations to know better their TIFF files for digital preservation.
• Using DPF Manager to check TIFF files.
5.4 Report on the PREFORMA hand-on seminar in Stockholm, May 29, 2017

On Monday the 29th of May 2017, the PREFORMA team at Riksarkivet (National Archives of Sweden) organised a so called “hands-on seminar” on how to validate compliance of files in PDF/A format with the standard specification and with the acceptance criteria of a memory institution (policy checking). The seminar was held in the premises of Riksarkivet in Stockholm and scheduled for a full day. The number of participants was by practical reasons limited to 25, but the invitation reached over 200 persons working with archives and records management issues at public and private memory institutions, state and communal agencies and organisations, and at SMEs (Small and Medium Sized Enterprises). 25 persons registered and was accepted for the seminar. Unfortunately, four of them could not attend because of illness or family reasons. Of the 21 that showed up:

- 12 came from state agencies and organisations, including 2 from public museums and archives and 3 from the offices of the Swedish Parliament and of the Swedish Government;
- 5 came from communal agencies, 3 from local and 2 from regional ones
- 3 came from SMEs targeting IT and archiving
- 1 came from a private archival institution.

After the participants had presented themselves and their expectations on this day, the seminar started with an introduction covering

- The aim of the seminar
- Presentation of the PREFORMA project
- An overview of the PREFORMA Challenge
  - What is Conformance checking?
  - What is Policy Checking?

Next item was a brief overview of the PDF/A format. All participants were to some extent familiar with the format, so this was more of an update and to equalize their knowledge.

After a coffee break, the hands-on part of the seminar started. All participants had to bring with them their own laptops (with Java and veraPDF version 1.4.6 installed) and examples of PDF/A files from their organisations. Technical support was given by the RA team to those who had problems in downloading required software or getting veraPDF to behave as expected.

The hands-on part was divided into the following sub-sessions:

- Conformance Checking in practice
  - Introduction
  - How to read reports and failure messages
  - How to identify and interpret failures
– Possible actions to solve failures

• Policy Checking in practice
  – Introduction
  – How to formulate a policy
  – How to test a policy

Each part of the sub-sessions started with a demonstration by the RA team and was then followed by practical exercises, where the participants used their own files or test files downloaded from Internet. Most of the participants also managed to create GitHub account and send in comments and questions.

During the day, a light lunch was served and of cause free coffee. Fruits were also on the table for the participants to consume whenever they need to boost their blood sugar level. The seminar ended with a discussion to evaluate the day. The expectations expressed in the beginning of the seminar as mainly focusing on learning more about PDF/A in order to better understand its “pros and cons” but also to learn about validation and the PREFORMA Conformance Checker. The feedback from the participants was very positive. Some commented that conformance checking requires knowledge about the format that most curators of digital object do not have today. Still confuse, but on a higher level, as one participant summarised it. More and better information on how to work in practice with the PREFORMA tools is therefore needed (tutorials was mentioned). More “fixers” was also asked for; the metadata fixer could be complemented with more simple fixing.

It was a consensus among the participants that PREFORMA is an important project and “spot on” so to speak, but some were anxious about the sustainability of the results and underlined the need for a follow-up project – in practice a PPI. It was also pointed out that an national institution in Sweden (in practice Riksarkivet) need to take the lead in implementing the PREFORMA tools and organise follow-up seminars on PDF/A and new seminars on TIFF and on MediaCronch.

All participants were interested to be part of a Swedish informal reference group.
5.5 Report on the PREFORMA hand-on seminar in Quedlinburg, May 29, 2017

This hands-on session was organized on the DPF-manager with the Museum Association of Saxony-Anhalt in Quedlinburg, May 29, 2017.

The Session was embedded in a general meeting of the Working Group Digitisation of the Museum-Association (AG Digitalisierung MVSA). There were 22 participants mainly from medium-sized and small museums comprising museum-directors, curators and some IT-people. The workshop lasted about 2 hours.

At the beginning, there was a general introduction into file-formats for digital preservation, especially for text and images and an introduction to illustrate PREFORMA’s goals and the tools created in the project.

Four participants brought a laptop and some TIFF-files along to be tested. All laptops were running on windows (different versions, 32 bit and 64 bit machines). There was no problem in downloading and installing the software, but many of the participants were convinced that the IT of the city or country that is responsible for their museum would have made troubles installing such software. This impression was not supported by any evidence, but it should be taken into account in order to simplify as much as possible the download and installation of the tools.

After the installation of the dpf-Manager, the functionalities available in the windows version were explained; the participants were able to understand “conformance”, “policy”, etc. Most functionalities worked smoothly. On Windows 10 (Touchscreen) there were problems with selecting files from pop-up menu “tree view” (scrolling in pop-up did not work). But otherwise everything worked fine.

There was some disappointment when the participants got the reports for their TIFF files: All were helpless, what to do and how to interpret the messages in purely technical English. What to do if a file was marked as not conforming? After some explanation we agreed that in any case it is good to know if a file is conform to the defined requirements or not.

At the end the participants agreed that the dpf-manager is a valuable tool for their digitization-work, not only for digital preservation but also to be used when checking image-files produced by external companies in the framework of a digitization project of the museum.

There is room for improvement: The messages should be in multilingual (German in this case) or at least there should be the possibility to set up languages other than English. It might be good to have some hints on possibilities (software tools) to correct found errors.

In general, it took some time to make clear, that conformance to baseline-TIFF is necessary for digital preservation. But after this was clarified, participants were eager to use the dpf-Manager tool. We are convinced that many of those who did not bring a laptop along will ask their IT-people to install the dpf-manager for them, others will test the tool at their private computer.

6 Feedback on the final release of the testing phase and on the EoP report

This section includes the comments provided by the members of the PREFORMA Evaluation Committee on the final release of the testing phase (end of August 2017) and on the End of Phase report.
6.1 Text Media Type

6.1.1 General comments.

Domain expert: It was a beneficial experience to test the veraPDF software as the National Library of Estonia is currently looking to develop a new digital archive which will need to have tools to validate its content. It’s good to test out software that we might use in the future.

6.1.2 The Conformance Checker.

Domain expert: We tested veraPDF with the print files that publishers have sent us. We converted them to PDF-A in order to test them. We couldn’t detect any issues when we compared the results with another validator but that might be because we didn’t test with enough files and also because we don’t have enough expertise yet when it comes to this tool.

Domain expert: Validates as it should. Can view reports in xml and html and save reports as it should. The reports could be more user friendly for non-technical users.

Domain expert: You are not given the option to choose profile, see Figure 17?

![Figure 17: Screen-shot of the VeraPDF Conformance Checker.](image_url)

Domain expert: When you choose policy, you are prompted to choose a file with the extension sch, xsl or xslt. It is not very clear to the user what they are required to do here.

Domain expert: You are offered a selection of report types to choose from, i.e. validation, features, validation and features and policy, however the report provided seems to be the same regardless of which option the users selects?
Domain expert: The conformance checker seems to be in acceptable state. The fast implementation if PDF/A 2.0 is a plus.

6.1.3 Result of compilation of CDP.

Technical expert:

- Offline compilation on Ubuntu 16.04 x86_64 desktop
- Cf previous build report in email: Re: [PREFORMA Evaluation Committee] veraPDF core distribution package released (2017-01-30 17:36:17)
- Still bundling Java from Oracle
  Bundled Java
  - java version "1.7.0_79"
  - Java(TM) SE Runtime Environment (build 1.7.0_79-b15)
  - Java HotSpot(TM) 64-Bit Server VM (build 24.79-b02, mixed mode)
  Bundled Maven
  - Apache Maven 3.3.9 (bb52d8502b132ec0a5a3f4c09453c07478323dc5; 2015-11-10T17:41:47+01:00)
  - Maven home: /opt/apache-maven-3.3.9
  - Java version: 1.7.0_79, vendor: Oracle Corporation
  - Java home: /usr/lib/jvm/oracle_jdk7/jre
  - Default locale: en_US, platform encoding: UTF-8
  - OS name: "linux", version: "4.4.0-31-generic", arch: "amd64", family: "unix"

Issues documented on https://github.com/veraPDF/veraPDF-library/issues/921

Successful compilation of CDP Implementation and Policy Checker (CLI):

- The compilation can be installed "as usual",
- some minor test of the implementation checker,
- the policy checker accepts policy schema and reports a result.

6.1.4 End of Phase Report.

The final report is very thorough and informative. It gives a good impression of the work done and of the results achieved. Here below is a short summary of the report, extrapolating the keywords and highlighting in bold some opinions and a few observations/notes made by the Evaluation Committee members.
Table 11: Summary of the report, extrapolating the keywords and highlighting in bold some opinions and a few observations/notes made by the Evaluation Committee members.

<table>
<thead>
<tr>
<th>Section</th>
<th>Report</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Details</td>
<td><strong>OK</strong></td>
</tr>
</tbody>
</table>
| 2. At the outset of this piece of work, what were your aims and objectives? | - utilise the testing results / automating testing  
- improve functionalities  
- work on specification / standardisation  
- work on documentation / support  
- community creating  
== Basically : Quality improvements |
| 3. Please provide a summary of the outputs of this piece of work and relate these to the original objectives. How do the outputs address the challenge of this PCP? | **PCP.** <- veraPDF test corpus, application, fixer  
- 1. <- real time testing  
- 2. <- OPF, DPC, KEEP  
- 3. <- PDF Validation TWG  
- 5. <- dissemination activities |
| 4. Describe any changes to the original plan in the tender. What was the reason for these changes? Please include any circumstances that aided or impeded the progress of the project and the actions taken to overcome them. | **GUI for creating Policy profiles**  
- improving the Feature report |
| 5. Please provide a short factual summary of the most significant outcomes of your work. | **automation of integrated tests**  
- widening scope of integrated tests  
- automated generation of PREFORMA testing Workflow  
- wide scale testing (European Publications Office)  
- test corpus review  
- new test corpus  
== testing breakthrough  
- improvements  
- support for PDF 2.0  
- documentation integration  
- influencing PDF/A-4  
- dissemination / community activity |
<p>| 6. Please explain how you addressed the recommendations provided during the previous phase. | <strong>documentation adjusted</strong> |</p>
<table>
<thead>
<tr>
<th>7. Describe the innovative aspects of the work, including any new findings or techniques.</th>
<th>unique network of stakeholders</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ISO WG collaboration</td>
</tr>
<tr>
<td></td>
<td>industry partnership</td>
</tr>
<tr>
<td></td>
<td>grammar based model and validation profiles</td>
</tr>
<tr>
<td></td>
<td>proven to be easily extendable (PDF 2.0)</td>
</tr>
<tr>
<td>8. Describe where the R&amp;D and other operational activities have been performed.</td>
<td>OK</td>
</tr>
<tr>
<td></td>
<td>OK</td>
</tr>
<tr>
<td>9. Please provide complete and clear information about the allocation of the money paid by the Authority taking into account the R&amp;D service contract minimum requirement (that more than 50% of the contract value is attributable directly and exclusively to legitimate R&amp;D services).</td>
<td></td>
</tr>
</tbody>
</table>
## 10. Datasets used to test the release
- Existing datasets
- Isartor
- BFO
- Digital corpora
- Newly created
- veraPDF corpus
- Memory institutions
- Open Source Community

## 11. Dissemination activities
- Web presence
- Events / conferences
- Articles
- Papers
- Blog (UX)
- Webinars
- Press releases
- “version 7 of callas pdfpilot supports verapdf”
- External press
- Twitter
- veraPDF news
- Mailing list / PDF Validation TWG / PDF Association

## 12. Level of activity generated in the open source community
- GitHub statistics
- Contributions / Issues
- Downloads
- StackOverflow presence

## 13. Describe any potential long-term collaboration / partnership entered into
- KEEP Solutions
- Artefactual Systems
- Logius
- The ZUGFeRD Community
- using veraPDF to validate PDF/A-3
- developing an open source veraPDF plugin to validate attached XML invoices

## 14. Standardisation efforts
- PDF Association’s PDF Validation TWG
- ISO TC 171 SC 2 WG 5
- Concrete influence on the development of next-generation PDF/A

## 15. Open Source approach
- GitHub
- Signed tags
- Travis-CI
### 16. Please describe what your organisation has gained from this project and what have been the main benefits. What new business opportunities have been created? Do you expect your organisation to grow as a result of this project?

- we have established strong relationships with consortium partners
- we have raised our profile
- we have advanced the cause of open source in digital preservation
- we have advanced the technology with the field improving the market
- we have identified the real problems and the methods for solving them
- we have new contacts and partnerships in the area of digital preservation
- we have provided the market with the tools to further develop the validation process

### 17. Describe the potential for exploiting the work. Please identify any new intellectual property which has been filed or for which filing is anticipated.

- See exploitation plan to be released in Sept 2017 (still pending)

### 18. Future plans

- See Roadmap and exploitation plan to be released in Sept 2017 (still pending)

### 19. Please add any additional information that you consider relevant to be reported. This may be in the form of text, pictures, diagrams, data, graphs that support the work done.

None

### 20. Describe what ethical aspects you have identified and how this may influence your project.

None

## 6.2 Image Media Type

### 6.2.1 General comments.

**Domain expert**: It will be used in our production environment soon, and as it is now, and of course we're looking forward to being able to get new features (policy checker related) when they are needed. The tool is suited for our requirements.
Domain & technical expert: I am trying this with the intention of using it in a hi-volume automated productions system so the GUI is less important for me as the command line is.

Domain expert: The DPF Manager worked well and fast enough it should be made part of digital preservation workflows in museums. But, aside from this, we discovered that it is also a useful tool for digitization projects. Often digitization in German museums is done by external companies, delivering TIFF – files. With DPF Manager the museums are enabled to check the quality of the work of these companies.

6.2.2 The Conformance Checker.

Domain expert: I’m very happy with the program and with the fast reaction when asking for some update (in my case, to be able to check a certain policy-feature which wasn’t available: they added it in the next release).

Domain expert: There's still the issue about using â˜œcmd-Hâ˜œI on the mac version. The program hides well as it should but when trying to return to the program using â˜œcmd-TABâ˜œI the program pops up (as it should) but the window doesn’t pop up, but it should.

Domain expert: They improve some of the tips that we made some months ago to manage better the addition of basic identification metadata and the statistic module.

Domain expert: The quick check is a new functionality that is much appreciated.

Domain & technical expert: I have tried DPF Manager for both Windows and Linux Ubuntu platforms up to version 3.5.1 of September-1 2017. The instructions for installing in non-GUI-Linux systems are still missing in the documentation, which means that if I am using a text-only system I am on my own.

Domain & technical expert: The software (both GUI and command line) should be made easier and more intuitive. Even I, as an expert user, find it hard to understand how to use the checker in a more detailed or specialized way, especially when I want to ignore errors in selected tags.

Domain expert: Installation of conformance checker was tested on a Debian standard system, everything went well. Trying to install it on a non-stable Debian version was not an easy task but we cannot blame the DPF Manager for this. One think we could not find – and that might also be out of scope of the DPF Manager – was an interface to provide easily translations of massages and reports. That is one possible point for further development beyond PREFORMA.

6.2.3 Result of compilation of CDP

Technical expert

- Offline compilation on Ubuntu 16.04 x86_64 desktop
- Cf previous build reported in email: Re: DPF Manager: issues to be addressed and next steps (2017-01-24 21:41:10).
- They have included the license files, but not bundled OpenJDK in the buildenv this time around (why...?).
Installed Ubuntu Java (default-jdk)

- openjdk version "1.8.0_131"
- OpenJDK Runtime Environment (build 1.8.0_131-8u131-b11-2ubuntu1.16.04.3-b11)
- OpenJDK 64-Bit Server VM (build 25.131-b11, mixed mode)

Bundled Maven

- Apache Maven 3.3.9 (bb52d8502b132ec0a5a3f4c09453c07478323dc5; 2015-11-10T17:41:47+01:00)
- Maven home: /media/user/buildenv/apache-maven-3.3.9
- Java version: 1.8.0_102, vendor: Azul Systems, Inc.
- Java home: /usr/lib/jvm/zulu-8-amd64/jre
- Default locale: en_US, platform encoding: UTF-8
- OS name: "linux", version: "4.4.0-31-generic", arch: "amd64", family: "unix"

Successful compilations of CDP Implementation and Policy Checker:

- tiffimplementationchecker-1.2.8-jar-with-dependencies.jar (tested a few files)
- tiffpolicychecker-1.1.6-jar-with-dependencies.jar (no test of policy but the program executes and prompts the help information)

6.2.4 End of Phase Report

The final report is very thorough and informative. It gives a good impression of the work done and of the results achieved. Here below is a short summary of the report, extrapolating the keywords and highlighting in bold some opinions and a few observations/notes made by the Evaluation Committee members.

<table>
<thead>
<tr>
<th>Section</th>
<th>Report</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Details</td>
<td>OK</td>
</tr>
</tbody>
</table>
| 2. At the outset of this piece of work, what were your aims and objectives? | • utilise the testing results  
• new functionalities / response to feedback  
• begin the standardisation process  
• community creating |

== Basically : Quality improvements
3. Please provide a summary of the outputs of this piece of work and relate these to the original objectives. How do the outputs address the challenge of this PCP?

<table>
<thead>
<tr>
<th>Outputs</th>
<th>Challenge Addressed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Test files &amp; instructions</td>
<td>1-2. Detailed reports, hints how to fix, statistical analysis, multiple running instances</td>
</tr>
<tr>
<td>2. TI/A submitted to ISO TC 171</td>
<td>3. Business plan</td>
</tr>
</tbody>
</table>

4. Describe any changes to the original plan in the tender. What was the reason for these changes? Please include any circumstances that aided or impeded the progress of the project and the actions taken to overcome them.

<table>
<thead>
<tr>
<th>Changes</th>
<th>Reason</th>
<th>Actions Taken</th>
</tr>
</thead>
<tbody>
<tr>
<td>75% test files</td>
<td>We had to create</td>
<td></td>
</tr>
<tr>
<td>Integration with Archivematica postponed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>New features: quick check</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5. Please provide a short factual summary of the most significant outcomes of your work.

<table>
<thead>
<tr>
<th>Outcomes</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instructions for creating testing files</td>
<td></td>
</tr>
<tr>
<td>New features (9x), improvements (30x)</td>
<td>Feedback</td>
</tr>
<tr>
<td>Most significant: statistical report module</td>
<td></td>
</tr>
<tr>
<td>They improve the reload button in statistics section</td>
<td></td>
</tr>
<tr>
<td>They improve the select/unselect all in policy checker invalidated rules</td>
<td></td>
</tr>
<tr>
<td>They improve the order reports by date and time</td>
<td></td>
</tr>
<tr>
<td>Possibility to copy text in report tooltips</td>
<td></td>
</tr>
<tr>
<td>Timer shows hours</td>
<td></td>
</tr>
<tr>
<td>Added hours to task time</td>
<td></td>
</tr>
<tr>
<td>Dissemination breakthrough</td>
<td></td>
</tr>
</tbody>
</table>

6. Please explain how you addressed the recommendations provided during the previous phase.

<table>
<thead>
<tr>
<th>Recommendations Addressed</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Documentation adjusted</td>
<td></td>
</tr>
<tr>
<td>Bugs fixed</td>
<td></td>
</tr>
<tr>
<td>UX improvements</td>
<td></td>
</tr>
<tr>
<td>Modification of implementation rules moved to policy</td>
<td></td>
</tr>
<tr>
<td>Using the &quot;Continuous Integration methodology&quot;</td>
<td></td>
</tr>
<tr>
<td>(was that a recommendation?)</td>
<td>Monthly builds delivered</td>
</tr>
<tr>
<td>CDP delivered</td>
<td></td>
</tr>
</tbody>
</table>
7. Describe the innovative aspects of the work, including any new findings or techniques.

- License fixes
- Roadmap/milestones update
- Unique network of stakeholders
- Technology – Maven, Travis, GitHub, modular architecture using event-driven communication, not-hard-coded/XML/rules engine
- Quick check
- UX – current trends / feedback
- TI/A

8. Describe where the R&D and other operational activities have been performed.

OK

9. Please provide complete and clear information about the allocation of the money paid by the Authority taking into account the R&D service contract minimum requirement (that more than 50% of the contract value is attributable directly and exclusively to legitimate R&D services).

OK

10. Datasets used to test the release

<table>
<thead>
<tr>
<th>datasets used to test the release</th>
</tr>
</thead>
<tbody>
<tr>
<td>- internal test files</td>
</tr>
<tr>
<td>- Ground truth test files</td>
</tr>
</tbody>
</table>

11. Dissemination activities

<table>
<thead>
<tr>
<th>dissemination activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>- TI/A external and internal website</td>
</tr>
<tr>
<td>- Twitter</td>
</tr>
<tr>
<td>- [image] scientific community / conferences</td>
</tr>
<tr>
<td>- courses</td>
</tr>
<tr>
<td>- collaboration (KOST-CECO)</td>
</tr>
<tr>
<td>- automatic feedback / early adopters</td>
</tr>
<tr>
<td>- newsletter / workshop / blogs</td>
</tr>
</tbody>
</table>

12. Level of activity generated in the open source community

<table>
<thead>
<tr>
<th>level of activity generated in the open source community</th>
</tr>
</thead>
<tbody>
<tr>
<td>- GitHub statistics</td>
</tr>
<tr>
<td>- “various institutions and companies are using our tool” – although not measured</td>
</tr>
<tr>
<td>- successful integration with Europeana D-Space project</td>
</tr>
</tbody>
</table>

13. Describe any potential long-term collaboration / partnership entered into

<table>
<thead>
<tr>
<th>describe any potential long-term collaboration / partnership entered into</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Digital humanities Lab (Basel, Switzerland)</td>
</tr>
<tr>
<td>- KOST (Switzerland)</td>
</tr>
<tr>
<td>- “Memory institutions”</td>
</tr>
<tr>
<td>- ISO TC-171</td>
</tr>
<tr>
<td>- Adobe Systems Inc (USA)</td>
</tr>
<tr>
<td>- AENOR (Spain)</td>
</tr>
<tr>
<td>- “TIFF Experts”</td>
</tr>
<tr>
<td>- Hewlett-Packard CDS (Spain)</td>
</tr>
<tr>
<td>14. Standardisation efforts</td>
</tr>
<tr>
<td>-----------------------------</td>
</tr>
</tbody>
</table>
| 15. Open Source approach    | • GPLv3++, MPLv2++  
|                             | • CDP, STD release  
|                             | • GitHub  
|                             | • roadmap / milestones  
|                             | • "project" feature  
|                             | • includes plans for TIFF-FX, TIFF-F, DNG  
|                             | • OpenJDK  
|                             | • "documentation concerning interpretation of technical specification of each file format"  
|                             | • "Redistributed in cascade"  |

| 16. Please describe what your organisation has gained from this project and what have been the main benefits. What new business opportunities have been created? Do you expect your organisation to grow as a result of this project? | • we have received valuable feedback  
|                                                                                   | • we have developed a unique specialised competence  
|                                                                                   | • we have become well-known within the field  
|                                                                                   | • we have established a large network  
|                                                                                   | • we have connected with key partners  
|                                                                                   | • we have become reputable within the field  
|                                                                                   | • we have identified a market  
|                                                                                   | • we are expanding  |

| 17. Describe the potential for exploiting the work. Please identify any new intellectual property which has been filed or for which filing is anticipated. | • Cloud-based/SaaS  
|                                                                                   | • On premise deployments  
|                                                                                   | • Technical support and maintenance contracts  
|                                                                                   | • Market place  
|                                                                                   | • Consultancy services  
|                                                                                   | • Training  
|                                                                                   | • Certification for service providers  |

| 18. Future plans | • Milestones defined into the 2020s  
|                 | • v 3.5.5 (Dec 2017)  
|                 | • v 3.6 (March 2018)  
|                 | • v 3.7 (Aug 2018)  
|                 | • v 3.8 (Dec 2018)  
|                 | • v 3.9 (June 2019)  
|                 | • v 4.0 (Nov 2019)  |
19. Please add any additional information that you consider relevant to be reported. This may be in the form of text, pictures, diagrams, data, graphs that support the work done.

20. Describe what ethical aspects you have identified and how this may influence your project.

<table>
<thead>
<tr>
<th>Projects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autofixes</td>
</tr>
<tr>
<td>Detect image transformations</td>
</tr>
<tr>
<td>XPath based implementation checker</td>
</tr>
<tr>
<td>TIFF-F, TIFF-FX</td>
</tr>
<tr>
<td>DNG</td>
</tr>
<tr>
<td>Advanced policy checker</td>
</tr>
</tbody>
</table>

OK

“... no relevant ethical aspects influenced our project.”

6.3 Audio/Video Media Type

6.3.1 General comments.

**External expert:** With implementing of the existing features of MediaInfo into MediaConch the field of use cases became wider than aimed by the project.

**Domain expert:** My one issue with MediaConch is that, to me, it is a tool for professionals. To work well for users that do not have a good technical knowledge it will have to be completed with, perhaps a knowledge base and a guide (guidance) on how to interpret conformance errors. As for now you get an error with a technical detail but no guidance on how this error could be interpreted or corrected. I think the software should benefit from adding a knowledge base to the software or the platform on Internet. The knowledge base should cover references to common errors and guidance on how to correct common errors. It could also include hints on how to deal with a more complex type of errors. This should probably make the software available to a broader public/group of users and increase its user base. In addition, an issue with the licensing is covered within the evaluation of the final report.

6.3.2 The Conformance Checker.

**External expert:** The fact that the standardization process will last longer than the project itself is a pity. As long as the standardization is ongoing, the conformance checking of the standard is no serious matter. However, the other areas of checking, such as the profiles (as presets or individual defined), still make sense.

**Domain expert:** The Conformance checker has been tested and it works quite well.

6.3.3 Result of compilation of CDP.

**Technical expert**

- Offline compilation on Ubuntu 16.04 x86_64 desktop
• Cf previous build reported in email: Re: Fwd: Re: MediaConch: issues to be addressed and next steps (2016-12-23 01:19:15).

Successful compilations of CDP Implementation Checker:
• CLI_compile.sh (tested a few files)
• GUI_compile.sh (tested a few files)
• Server_compile.sh (no test of files but the program executes and prompts the help information)

Successful compilations of CDP Policy Checker:
• CLI_compile.sh (no test of policy but the program executes and prompts the help information)
• GUI_compile.sh (no test of policy but the program executes)
• Server_compile.sh (no test of policy but the program executes and prompts the help information)

6.3.4 End of Phase Report

The final report is very thorough and informative. It gives a good impression of the work done and of the results achieved. It also shows the will of the developer to continue to improve the conformance checker after the end of the project. Working with standardization bodies (e.g. IETF CELLAR) is a good way to move forward. Here below is a short summary of the report, extrapolating the keywords and highlighting in bold some opinions and a few observations/notes made by the Evaluation Committee members.

Table 13: Summary of the report, extrapolating the keywords and highlighting in bold some opinions and a few observations/notes made by the Evaluation Committee members.

<table>
<thead>
<tr>
<th>Section</th>
<th>Report</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Details</td>
<td>OK</td>
</tr>
<tr>
<td>2. At the outset of this piece of work, what were your aims and objectives?</td>
<td>• utilise the testing results / automating testing</td>
</tr>
<tr>
<td></td>
<td>• new functionalities / response to feedback</td>
</tr>
<tr>
<td></td>
<td>• work on specification / standardisation</td>
</tr>
<tr>
<td>== Basically : Quality improvements</td>
<td></td>
</tr>
<tr>
<td>3. Please provide a summary of the outputs of this piece of work and relate these to the original objectives. How do the outputs address the challenge of this PCP?</td>
<td>• 1. &lt;- many test files</td>
</tr>
<tr>
<td></td>
<td>• 2. &lt;- stable software with rich functionalities, integration with Archivematica done</td>
</tr>
<tr>
<td></td>
<td>• 3. &lt;- IETF CELLAR</td>
</tr>
<tr>
<td>4. Describe any changes to the original plan in the tender. What was the reason for these changes? Please include any circumstances that aided or impeded the progress of the project and the actions taken to overcome them.</td>
<td>• integrating tests into an automated testing suite</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>
| 5. Please provide a short factual summary of the most significant outcomes of your work. | • dissemination / community breakthrough  
• wider platform support / distribution  
• More information included in the Dissemination Activities sections of the progress reports. |
| 6. Please explain how you addressed the recommendations provided during the previous phase. | • documentation adjusted  
• Several ‘fixer’ functions have been integrated into MediaConch according to user priority. The most notable one is where the CRCs of FFV1 and Matroska are used to repair some limited instances of digital damage. There’s an article about the fixer at https://mediaarea.net/MediaConch/fixity.html. There are also fixes for elemental size and structural issues, such as when a master element is labelled as having a zero size but followed by a child element of that master element.  
• UX improvements  
• large case projects: Archivematica + interview series (see q15 for open source) |
| 7. Describe the innovative aspects of the work, including any new findings or techniques. | • IETF CELLAR  
• "new insight into existing implementations of Matroska"  
• benchmark comparison (MXF, JP2K) gave new insights  
• policy checker |
| 8. Describe where the R&D and other operational activities have been performed. | OK / no physical office |
9. Please provide complete and clear information about the allocation of the money paid by the Authority taking into account the R&D service contract minimum requirement (that more than 50% of the contract value is attributable directly and exclusively to legitimate R&D services).

| OK / lightweight / minimal overhead |

<table>
<thead>
<tr>
<th>datasets used to test the release</th>
</tr>
</thead>
<tbody>
<tr>
<td>IETF CELLAR WG:s EBML Schema for Matroska</td>
</tr>
<tr>
<td>Archive.org corpus</td>
</tr>
<tr>
<td>Ground truth test files</td>
</tr>
</tbody>
</table>

10. Datasets used to test the release

<table>
<thead>
<tr>
<th>dissemination activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>distribution (package repositories)</td>
</tr>
<tr>
<td>seminar / workshop</td>
</tr>
<tr>
<td>collaboration (VIAA)</td>
</tr>
<tr>
<td>conferences</td>
</tr>
<tr>
<td>blog series</td>
</tr>
<tr>
<td>newsletter</td>
</tr>
<tr>
<td>As already said, we would have liked to see more information on targeted community building activities.</td>
</tr>
</tbody>
</table>

11. Dissemination activities

<table>
<thead>
<tr>
<th>level of activity generated in the open source community</th>
</tr>
</thead>
<tbody>
<tr>
<td>GitHub statistics</td>
</tr>
<tr>
<td>open specification process</td>
</tr>
<tr>
<td>user feedback; 1-on-1 coaching</td>
</tr>
</tbody>
</table>

12. Level of activity generated in the open source community

<table>
<thead>
<tr>
<th>describe any potential long-term collaboration / partnership entered into</th>
</tr>
</thead>
<tbody>
<tr>
<td>Artefactual (Canada)</td>
</tr>
<tr>
<td>Indiana University (USA)</td>
</tr>
<tr>
<td>IETF (international)</td>
</tr>
<tr>
<td>Library of Congress (USA)</td>
</tr>
<tr>
<td>PACKED</td>
</tr>
<tr>
<td>Tate Modern (UK)</td>
</tr>
<tr>
<td>VIAA</td>
</tr>
<tr>
<td>interviewed: Austrian Mediathek (Austria), Carnegie Hall (USA), Denver Art Museum (USA), Irish Film Archive (USA)</td>
</tr>
<tr>
<td>Confirmed MediaConch users: Indiana University (USA), National A/V Archive (Luxembourg), New York Public Library (USA), Reto Kromer (Swiss/Global), Smithsonian Folkways, LoC (USA), Tate (UK), University of North Carolina (USA)</td>
</tr>
</tbody>
</table>

13. Describe any potential long-term collaboration / partnership entered into

<table>
<thead>
<tr>
<th>standardisation efforts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extensive work have been done</td>
</tr>
</tbody>
</table>
15. Open Source approach

- FLAC
- license requirements for sqlite and libpng
- CDP, STD
- more work needs to be done. From the licenses file: WebUI engine (optional): Symfony, MIT license, could be relicensed to GPLv3 or later / MPLv2 or later, WebUI design (optional): Bootstrap, MIT license, could be relicensed to GPLv3 or later / MPLv2 or later, XML parser (optional): TinyXML-2, zlib license, could be relicensed to GPLv3 or later / MPLv2 or later, gzip decompression routine (optional): zlib, zlib license, could be relicensed to GPLv3 or later / MPLv2 or later, FTP, FTPS, SFTP, HTTP, HTTPS (optional): curl, MIT license, could be relicensed to GPLv3 or later / MPLv2 or later, Base64 (optional): base64 by Bob Withers, public domain, could be relicensed to GPLv3 or later / MPLv2 or later, MD5 (optional): md5 by Colin Plumb, public domain, could be relicensed to GPLv3 or later / MPLv2 or later

16. Please describe what your organisation has gained from this project and what have been the main benefits. What new business opportunities have been created? Do you expect your organisation to grow as a result of this project?

- we have gained experience and actual results to market
- we have expanded
- we have developed strong business links
- demand – supply required funding – demand not willing to take the risk/could not afford the costs – PREFORMA funding allowed the initial development
- we have connected with sponsors
- we have started an initiative to promote standardization of open formats; OLLISTD (Open LossLess in STanDards)

17. Describe the potential for exploiting the work. Please identify any new intellectual property which has been filed or for which filing is anticipated.

- requesting new features
- file format identification and validation
- standardizing or mapping file formats
- existing system integration
| 18. Future plans | • new public releases every 2 months (features depend on sponsors)  
|                | • LTS (2018)  
|                | • Standardization process |

| 19. Please add any additional information that you consider relevant to be reported. This may be in the form of text, pictures, diagrams, data, graphs that support the work done. | OK |

| 20. Describe what ethical aspects you have identified and how this may influence your project. | • Identified several issues: the licensing issues, "... objecting to a coordinator and collaborating with competitors...", "... to understand the tactics and philosophies of the organizations.", and how it affected the MediaArea's decision to organize team members. |
A Classes

A.1 Text Media Type

- Class TC000 – Correct
  - description: This class indicates the documents that do not have any conformance and/or policy issue.
  - PDF/A version: All
  - conformance level: All
  - class type: Conformance
  - severity: 0

- Class TC002 – Annotation FileAttachment non-PDF/A
  - description: Has Attachment of ANY kind of resource EXCEPT PDF/A-1 and -2.
  - PDF/A version: 2
  - conformance level: bua
  - class type: Conformance
  - severity: 5

- Class TC003 – Annotation Sound
  - description: Has Annotation using sound.
  - PDF/A version: 1-3
  - conformance level: bua
  - class type: Conformance
  - severity: 5

- Class TC004 – Annotation Movie
  - description: Has Annotation using movie.
  - PDF/A version: 1-3
  - conformance level: bua
  - class type: Conformance
  - severity: 5

- Class TC005 – Annotation Screen
  - description: Has Area on page where media can be played or enable other activites.
  - PDF/A version: 1-3
• conformance level: bua
  • class type: Conformance
  • severity: 5

- Class TC006 – Annotation 3D
  - description: Has 3D objects.
  - PDF/A version: 1-3
  - conformance level: bua
  - class type: Conformance
  - severity: 5

- Class TC007 – Encoding LZW
  - description: Has any information encoded using LZW compression.
  - PDF/A version: 1-3
  - conformance level: bua
  - class type: Conformance
  - severity: 5

- Class TC008 – Encoding Crypt
  - description: Has any information encoded using Crypt encryption.
  - PDF/A version: 1
  - conformance level: ba
  - class type: Conformance
  - severity: 5

- Class TC009 – Encoding Crypt
  - description: Has any information encoded using Crypt encryption with non-Identity decode parameter.
  - PDF/A version: 2-3
  - conformance level: bua
  - class type: Conformance
  - severity: 5

- Class TC010 – Image Encoding Interpolation
  - description: Has image encoded with Interpolation.
  - PDF/A version: 1-3
– conformance level: bua
– class type: Conformance
– severity: 5

• Class TC012 – **Document Optional content**
  – description: Has optional content ("layers").
  – PDF/A version: 1
  – conformance level: ba
  – class type: Conformance
  – severity: 5

• Class TC013 – **Transitions**
  – description: Has effects when, e.g., transitioning from one slide to another.
  – PDF/A version: 1-3
  – conformance level: bua
  – class type: Conformance
  – severity: 5

• Class TC014 – **Transparency**
  – description: Has PDF transparency.
  – PDF/A version: 1
  – conformance level: ba
  – class type: Conformance
  – severity: 5

• Class TC016 – **Document Attachment**
  – description: Has Attachment of ANY kind of resource.
  – PDF/A version: 1
  – conformance level: ba
  – class type: Conformance
  – severity: 5

• Class TC017 – **Document non-PDF/A Attachment**
  – description: Has Attachment of ANY kind of resource EXCEPT PDF/A-1 and -2.
  – PDF/A version: 2
  – conformance level: bua
– class type: Conformance
– severity: 5

• Class TC018 – **Executable PostScript**

  – description: Has executable PostScript.
  – PDF/A version: 1-3
  – conformance level: bua
  – class type: Conformance
  – severity: 5

• Class TC019 – **Form Action**

  – description: Has action specific to Forms.
  – PDF/A version: 1-3
  – conformance level: bua
  – class type: Conformance
  – severity: 5

• Class TC020 – **XFA Forms**

  – description: Has XML-based forms.
  – PDF/A version: 1-3
  – conformance level: bua
  – class type: Conformance
  – severity: 5

• Class TC021 – **Action Launch**

  – description: Has action to open a file or execute a program.
  – PDF/A version: 1-3
  – conformance level: bua
  – class type: Conformance
  – severity: 5

• Class TC022 – **Action Sound**

  – description: Has action to play a sound.
  – PDF/A version: 1-3
  – conformance level: bua
  – class type: Conformance
- severity: 5

• Class TC023 – **Action Movie**
  - description: Has action to play a movie
  - PDF/A version: 1-3
  - conformance level: bua
  - class type: Conformance
  - severity: 5

• Class TC024 – **Action Hide**
  - description: Has action to hide annotations or outlines. Actions are associated with annotations (including interactive forms) or outlines (bookmarks).
  - PDF/A version: 1-3
  - conformance level: bua
  - class type: Conformance
  - severity: 5

• Class TC025 – **Action ResetForm**
  - description: Has action to reset form, that is, clear the form of any input.
  - PDF/A version: 1-3
  - conformance level: bua
  - class type: Conformance
  - severity: 5

• Class TC026 – **Action ImportData**
  - description: Has action to import data. A conforming processor shall import Forms Data Format (FDF) data into the document's interactive form from a specified file (PDF reference, sec. 12.7.5.4).
  - PDF/A version: 1-3
  - conformance level: bua
  - class type: Conformance
  - severity: 5

• Class TC027 – **Action Javascript**
  - description: Has action to invoke JavaScript. A conforming processor shall execute a script that is written in the JavaScript programming language (PDF reference, sec. 12.6.4.16).
- **PDF/A version**: 1-3
- **conformance level**: bua
- **class type**: Conformance
- **severity**: 5

- **Class TC028 – Action Set-state**
  - **description**: Has the set-state action. Is obsolete and should not be used (PDF reference "NOTE", p. 418).
  - **PDF/A version**: 1-3
  - **conformance level**: bua
  - **class type**: Conformance
  - **severity**: 5

- **Class TC029 – Action No-op**
  - **description**: Has the No-op action. Obsolete.
  - **PDF/A version**: 1-3
  - **conformance level**: bua
  - **class type**: Conformance
  - **severity**: 5

- **Class TC030 – Action SetOCGState**
  - **description**: Has a set-OCG-state action. Sets the state of one or more optional content groups (PDF reference, sec. 12.6.4.12). Introduced in PDF 1.5.
  - **PDF/A version**: 1-3
  - **conformance level**: bua
  - **class type**: Conformance
  - **severity**: 5

- **Class TC031 – Action Rendition**
  - **description**: Has a rendition action. Controls the playing of multimedia content (PDF reference, sec. 12.6.4.12). Introduced in PDF 1.5.
  - **PDF/A version**: 1-3
  - **conformance level**: bua
  - **class type**: Conformance
  - **severity**: 5
• Class TC032 – Action Trans
  – description: Has a transition action. May be used to control drawing during a sequence of actions (PDF reference, sec. 12.6.4.14). Introduced in PDF 1.5.
  – PDF/A version: 1-3
  – conformance level: bua
  – class type: Conformance
  – severity: 5

• Class TC033 – Action GoTo3DView
  – PDF/A version: 1-3
  – conformance level: bua
  – class type: Conformance
  – severity: 5

• Class TC034 – Named Action
  – PDF/A version: 1-3
  – conformance level: bua
  – class type: Conformance
  – severity: 5

• Class TC035 – Encryption
  – description: Has general encryption (PDF reference, sec. 7.6).
  – PDF/A version: 1-3
  – conformance level: bua
  – class type: Conformance
  – severity: 5

• Class TC036 – Permission
  – description: Has user rights to document
  – PDF/A version: 2-3
  – conformance level: bua
- class type: Conformance
- severity: 5

- Class TC037 – .notdef
  - description: Has the .notdef glyph.
  - PDF/A version: 2-3
  - conformance level: bua
  - class type: Conformance
  - severity: 5

- Class TC038 – File header
  - description: Has
  - PDF/A version: 1
  - conformance level: bua
  - class type: Conformance
  - severity: 5

- Class TC039 – File header
  - description: Has
  - PDF/A version: 2-3
  - conformance level: bua
  - class type: Conformance
  - severity: 5

- Class TC040 – Keyword spacings
  - PDF/A version: 1-3
  - conformance level: bua
  - class type: Conformance
  - severity: 5

- Class TC041 – Does not have: Color space
  - description: Does not have a defined color space. PDF reference, sec. 8.6.
  - PDF/A version: 1-3
  - conformance level: bua
  - class type: Conformance
- severity: 5

• Class TC042 – Device dependent
  
  - description: Does not have a Device dependent color space. The device colour spaces (DeviceCMYK, DeviceGray, DeviceRGB) enable a page description to specify colour values that are directly related to their representation on an output device (see PDF reference, sec. 8.6.4.1).
  
  - PDF/A version: 1-3
  
  - conformance level: bua
  
  - class type: Conformance
  
  - severity: 5

• Class TC043 – Logical structure
  
  - description: Does not have structured content (see PDF reference, sec. 14.7).
  
  - PDF/A version: 1-3
  
  - conformance level: a
  
  - class type: Conformance
  
  - severity: 5

• Class TC044 – RoleMap
  
  - description: Has custom tags but without mapping them to the standard tags.
  
  - PDF/A version: 1-3
  
  - conformance level: a
  
  - class type: Conformance
  
  - severity: 5

• Class TC045 – Hierarchy
  
  - description: Does not have the logical structure of the document described by a hierarchy of objects called the structure hierarchy or structure tree (see PDF reference 14.7.2).
  
  - PDF/A version: 1-3
  
  - conformance level: a
  
  - class type: Conformance
  
  - severity: 5

• Class TC046 – Structure type
- **description**: Does not have structure type for every structure element. A name object that identifies the nature of the structure element and its role within the document (such as a chapter, paragraph, or footnote) (see PDF reference, sec. 14.7.3). This covers Tagged PDF (PDF 1.4), which is a stylized use of PDF that builds on the logical structure framework described in 14.7, “Logical Structure.” It defines a set of standard structure types and attributes that allow page content (text, graphics, and images) to be extracted and reused for other purposes (see PDF reference 14.8).
  - **PDF/A version**: 1-3
  - **conformance level**: a
  - **class type**: Conformance
  - **severity**: 5

- **Class TC047** – **Metadata (DocumentInfo)**
  - **description**: Does not have the document metadata in sync with the XMP metadata.
  - **PDF/A version**: 1
  - **conformance level**: ba
  - **class type**: Conformance
  - **severity**: 5

- **Class TC048** – **XMP (metadata)**
  - **description**: Does not have the document Metadata key XMP present.
  - **PDF/A version**: 1-3
  - **conformance level**: bua
  - **class type**: Conformance
  - **severity**: 5

- **Class TC049** – **Undefined XMP properties**
  - **description**: Has XMP metadata with custom properties but without defining them in a schema.
  - **PDF/A version**: 1-3
  - **conformance level**: bua
  - **class type**: Conformance
  - **severity**: 5

- **Class TC050** – **XMP PDF/A version identifier**
  - **description**: Has XMP metadata but the presence of the PDF/A version or conformance element is missing.
– **PDF/A version**: 1-3
– **conformance level**: bua
– **class type**: Conformance
– **severity**: 5

- **Class TC051 – `<pdfaid:part>`<1-3>`**
  – **description**: Has XMP metadata AND PDF/A version element BUT not in correct format.
  – **PDF/A version**: 1-3
  – **conformance level**: bua
  – **class type**: Conformance
  – **severity**: 5

- **Class TC052 – `<pdfaid:conformance>`<BUA>`**
  – **description**: Has XMP metadata AND PDF/A conformance element BUT not in correct format.
  – **PDF/A version**: 1-3
  – **conformance level**: bua
  – **class type**: Conformance
  – **severity**: 5

- **Class TC053 – Embed composite fonts CMaps**
  – **description**: Does not have defined mappings from Unicode enodings to character collections (PDF reference, sec. 9.7.5.2).
  – **PDF/A version**: 1
  – **conformance level**: ba
  – **class type**: Conformance
  – **severity**: 5

- **Class TC054 – Embed non-predefined CMaps**
  – **description**: Does not have defined mappings from Unicode enodings to character collections (PDF reference, sec. 9.7.5.2).
  – **PDF/A version**: 2-3
  – **conformance level**: bua
  – **class type**: Conformance
  – **severity**: 5
• Class *TC055* – **Embed fonts**
  - *description*: Does not have fonts embedded (PDF reference, sec. 9.9).
  - *PDF/A version*: 1-3
  - *conformance level*: bua
  - *class type*: Conformance
  - *severity*: 5

• Class *TC056* – **Embed fonts Text rendering mode 3**
  - *description*: Does not have fonts not rendered embedded.
  - *PDF/A version*: 1-3
  - *conformance level*: bua
  - *class type*: Conformance
  - *severity*: 5

• Class *TC058* – **Character identification of font subsets**
  - *description*: Does not have Keys CharSet and CIDSet.
  - *PDF/A version*: 1-3
  - *conformance level*: bua
  - *class type*: Conformance
  - *severity*: 5

• Class *TC059* – **Unicode Character map**
  - *description*: Does not have characters mapped to Unicode table (PDF reference, sec. 9.10.3).
  - *PDF/A version*: 1-3
  - *conformance level*: ua
  - *class type*: Conformance
  - *severity*: 5

• Class *TC060* – **Font metric**
  - *PDF/A version*: 1-3
  - *conformance level*: bua
  - *class type*: Conformance
  - *severity*: 5
- **Class TC062 – Data after EOF**
  - *description*: Has data after the EOF marker.
  - *PDF/A version*: 1-3
  - *conformance level*: bua
  - *class type*: Conformance
  - *severity*: 5

- **Class TC064 – Conformance Level U**
  - *description*: Does not have functionality required for Conformance Level U.
  - *PDF/A version*: 2-3
  - *conformance level*: u
  - *class type*: Conformance
  - *severity*: 5

- **Class TC065 – Conformance Level A**
  - *description*: Does not have functionality required for Conformance Level A.
  - *PDF/A version*: 1-3
  - *conformance level*: a
  - *class type*: Conformance
  - *severity*: 5

- **Class TC066 – Specified font**
  - *description*: Does not have specified font.
  - *PDF/A version*: Any
  - *conformance level*: Any
  - *class type*: Policy
  - *severity*: 5

- **Class TC067 – Unspecified font**
  - *description*: Does have a disallowed font
  - *PDF/A version*: Any
  - *conformance level*: Any
  - *class type*: Policy
  - *severity*: 5

- **Class TC068 – JPG (Codec) is used in the document.**
- **description**: Does not have a JPG codec in document.
- **PDF/A version**: Any
- **conformance level**: Any
- **class type**: Policy
- **severity**: 5

- **Class TC069 – JPX (Codec) is used in the document.**
  - **description**: Does not have JPX codec in document.
  - **PDF/A version**: Any
  - **conformance level**: Any
  - **class type**: Policy
  - **severity**: 5

- **Class TC070 – CCITT (Codec) is used in the document**
  - **description**: Has CCITT codec in document.
  - **PDF/A version**: Any
  - **conformance level**: Any
  - **class type**: Policy
  - **severity**: 5

- **Class TC071 – LZW is used in the document**
  - **description**: Does not have LZW codec.
  - **PDF/A version**: Any
  - **conformance level**: Any
  - **class type**: Policy
  - **severity**: 5

- **Class TC073 – Author property in metadata is absent or empty**
  - **description**: Does not have text in XMP author property and DocumentInfo.
  - **PDF/A version**: Any
  - **conformance level**: Any
  - **class type**: Policy
  - **severity**: 5

- **Class TC074 – Does have: a flag Copyright protected fonts**
– *description*: Has a flag for Copyright protected fonts. The embedded font contains a permissions flag specifying that the font is not allowed for embedding without a special permission from the copyright holder
  – *PDF/A version*: Any
  – *conformance level*: Any
  – *class type*: Policy
  – *severity*: 5

• **Class TC075 – Embedded audio/video**
  – *description*: Does not have audio/video content.
  – *PDF/A version*: Any
  – *conformance level*: Any
  – *class type*: Policy
  – *severity*: 5

• **Class TC077 – Encryption is allowed**
  – *description*: The document is not encrypted.
  – *PDF/A version*: Any
  – *conformance level*: Any
  – *class type*: Policy
  – *severity*: 5

• **Class TC078 – Specified title property in metadata**
  – *description*: The document does not have title property in metadata
  – *PDF/A version*: Any
  – *conformance level*: Any
  – *class type*: Policy
  – *severity*: 5

• **Class TC079 – Each page is a single image**
  – *description*: Each page IS a single image, suggesting a scanned document
  – *PDF/A version*: Any
  – *conformance level*: Any
  – *class type*: Policy
  – *severity*: 5
• Class TC080 – **Each page is a single bitonal image**
  
  – *description*: Each page IS a single bitonal image, suggesting a scanned text document
  – *PDF/A version*: Any
  – *conformance level*: Any
  – *class type*: Policy
  – *severity*: 5

• Class TC081 – **Each page is a single colour image**
  
  – *description*: Each page IS a single colour image
  – *PDF/A version*: Any
  – *conformance level*: Any
  – *class type*: Policy
  – *severity*: 5

• Class TC082 – **Specified document structure tree**
  
  – *description*: The document does not have a structure tree
  – *PDF/A version*: Any
  – *conformance level*: Any
  – *class type*: Policy
  – *severity*: 5

• Class TC083 – **Specified language**
  
  – *description*: The document does not contain specified language
  – *PDF/A version*: Any
  – *conformance level*: Any
  – *class type*: Policy
  – *severity*: 5

• Class TC084 – **Unspecified languages**
  
  – *description*: The document does contain unspecified languages (there may be many in a document) outside of a restricted set
  – *PDF/A version*: Any
  – *conformance level*: Any
  – *class type*: Policy
  – *severity*: 5
• Class **TC085 – Specified PDF producer**
  
  - *description*: The document producer is not the specified producer
  - *PDF/A version*: Any
  - *conformance level*: Any
  - *class type*: Policy
  - *severity*: 5

• Class **TC086 – Specified PDF version (1.5, 1.5, 1.6, 1.7)**
  
  - *description*: The PDF version is not the specified value
  - *PDF/A version*: Any
  - *conformance level*: Any
  - *class type*: Policy
  - *severity*: 5

### A.2 Image Media Type

• Class **IC000 – Correct**
  
  - *description*: This class indicates the documents that do not have any conformance and/or policy issue.
  - *class type*: Conformance
  - *severity*: 0

• Class **IC001 – BigTiff**
  
  - *description*: TIFF with 64 bit offsets
  - *class type*: Conformance
  - *severity*: 5

• Class **IC002 – Tag cardinality**
  
  - *description*: TIFF using a Tag with incorrect cardinality (Other cardinality than specified in Baseline 6.0)
  - *class type*: Conformance
  - *severity*: 5

• Class **IC003 – Incorrect tag type**
  
  - *description*: TIFF with a Tag with incorrect type but still readable (TIFF readers should accept BYTE, SHORT, or LONG values for any unsigned integer field. [Section 2, page 15]
- **class type**: Conformance
- **severity**: 4

• **Class IC004 – Incompatible tag type**
  - **description**: TIFF using a Tag with incorrect and incompatible type (Other type than specified in Baseline 6.0)
  - **class type**: Conformance
  - **severity**: 5

• **Class IC005 – Channels error**
  - **description**: Channels count do not match (SamplesPerPixel, ExtraSamples and BitsPerPixelSample must have consistency)
  - **class type**: Conformance
  - **severity**: 5

• **Class IC006 – Dimensions error**
  - **description**: Incorrect image width/height (the Image width or Height declared not match with the tiled or striped image data structure)
  - **class type**: Conformance
  - **severity**: 5

• **Class IC007 – Resolution error**
  - **description**: Missing Resolution (XResolution and YResolution tag) or declared with a zero value [Section 7, page 27]
  - **class type**: Conformance
  - **severity**: 3

• **Class IC008 – Missing required tags in an Image IFD Baseline 6.0**
  - **description**: Required tags for defining an Image IFD not present (ImageWidth, ImageLength, XResolution, YResolution, PhotometricInterpretation)
  - **class type**: Conformance
  - **severity**: 5

• **Class IC009 – Incorrect page number**
  - **description**: A TIFF multipage document (NewSubFileTyp values 2,3,6 or 7) with incorrect page number (page numbers must range from zero to the number of images, missing pages, duplicat pages, inconsistent number of pages) [Section 12, page 55]
– class type: Conformance
– severity: 1

• Class **IC010** – **Unexpected tag type**
  – description: Unknown tag type (type not defined in Baseline 6.0)
  – class type: Conformance
  – severity: 3

• Class **IC011** – **MagicNumber**
  – description: Tiff signature not correct (magic number must be 42) [Section 2, page 13]
  – class type: Conformance
  – severity: 5

• Class **IC012** – **Byte Order**
  – description: Incorrect Byte Order (only little endian or big endian are accepted) [Section 2, page 13]
  – class type: Conformance
  – severity: 5

• Class **IC013** – **Bad alignment**
  – description: Offsets in a word boundary [Section 2, page 13]
  – class type: Conformance
  – severity: 4

• Class **IC014** – **Bad Offset**
  – description: IFD or tags offsets pointing outside the file or inside the Image File Header, re-use offsets (offset points or overlapping to already used data) [Section 2, page 14]
  – class type: Conformance
  – severity: 5

• Class **IC015** – **IFD entries 0**
  – description: Number of IDF entries in an IFD is zero [Section 2, page 14]
  – class type: Conformance
  – severity: 5

• Class **IC016** – **No IFDs**
  – description: At least 1 IFD must exist [Section 2, page 14]
- **class type:** Conformance
- **severity:** 5

- **Class IC017 – IFD Entry not in ascending order**
  - **description:** Tags not ordered in strict ascending order (tags not ordered or duplicate tags) [Section 2, page 15]
  - **class type:** Conformance
  - **severity:** 4

- **Class IC018 – Invalid Photometric Interpretation**
  - **description:** Photometric Interpretation must be a valid value defined in the TIFF Baseline 6.0 [Section 8, page 27]
  - **class type:** Conformance
  - **severity:** 5

- **Class IC019 – Coherent strips tags**
  - **description:** StripOffsets and StripByteCount cardinalities matching [Section 8, page 40]
  - **class type:** Conformance
  - **severity:** 5

- **Class IC020 – Strips tags**
  - **description:** StripOffsets, StripByteCounts tags must exist for stripped images [Section 8, page 40]
  - **class type:** Conformance
  - **severity:** 5

- **Class IC021 – Consistent strips**
  - **description:** Check strips sizes match image dimensions [Section 8, page 40]
  - **class type:** Conformance
  - **severity:** 5

- **Class IC022 – Coherent tiles tags**
  - **description:** TilesOffsets and TileByteCount cardinalities matching [Section 15, page 66]
  - **class type:** Conformance
  - **severity:** 5

- **Class IC023 – Tiles tags**
– **description**: TileOffsets, TileWidth, TileByteCounts and TileLength tags must exist for tiled images [Section 15, page 66]
  – **class type**: Conformance
  – **severity**: 5

- **Class IC024 – Valid tile tags**
  – **description**: TileWidth, and TileLength greater than zero [Section 15, page 66]
  – **class type**: Conformance
  – **severity**: 5

- **Class IC025 – Consistent tiles**
  – **description**: Check tiles sizes match image dimensions [Section 15, page 66]
  – **class type**: Conformance
  – **severity**: 5

- **Class IC026 – Bilevel**
  – **description**: Incorrect tags for Bilevel images [Section 3, page 21]
  – **class type**: Conformance
  – **severity**: 5

- **Class IC027 – Grayscale**
  – **description**: Incorrect tags for Grayscale images [Section 4, page 22]
  – **class type**: Conformance
  – **severity**: 5

- **Class IC028 – Pallete**
  – **description**: Incorrect tags for Pallete images (ColorMap and InkSet are mandatory) [Section 5, page 23]
  – **class type**: Conformance
  – **severity**: 5

- **Class IC029 – Transparency Mask**
  – **description**: Incorrect tags for Transparency mask images [Section 8, page 37]
  – **class type**: Conformance
  – **severity**: 5

- **Class IC030 – CMYK**
• **description:** Incorrect tags for CMYK images [Section 16, page 69]
  • **class type:** Conformance
  • **severity:** 5

• Class **IC031 – YCbCr**
  • **description:** Incorrect tags for YCbCr images [Section 21, page 94]
  • **class type:** Conformance
  • **severity:** 5

• Class **IC032 – CIELab**
  • **description:** Incorrect tags for CIELab images [Section 23, page 110]
  • **class type:** Conformance
  • **severity:** 5

• Class **IC033 – RGB**
  • **description:** Incorrect tags for RGB images [Section 6, page 24]
  • **class type:** Conformance
  • **severity:** 5

• Class **IC034 – Bad Ascii7 format**
  • **description:** Tags with Ascii format containing non-7 bits ascii, Ascii without null character termination, More than one null between strings [Section 2, page 15]
  • **class type:** Conformance
  • **severity:** 3

• Class **IC035 – Bad Datetime**
  • **description:** Datetime tag with incorrect format, invalid date, incorrect tag type (No ASCII), incorrect cardinality [Section 8, page 31]
  • **class type:** Conformance
  • **severity:** 3

• Class **IC036 – Private tags**
  • **description:** If more than 10 private tags are used in one IFD, a private IFD should be used to encapsulate them [Section 0, page 9]
  • **class type:** Conformance
  • **severity:** 1
• Class IC037 – TiffEP
  – description: Incorrect Tiff EP
  – class type: Conformance
  – severity: 1

• Class IC038 – TiffEP StandardID
  – description: Incorrect Tiff EP with tag TIFF/EPStandardID
  – class type: Conformance
  – severity: 5

• Class IC039 – TiffIT
  – description: Incorrect Tiff IT
  – class type: Conformance
  – severity: 1

• Class IC040 – Lossy compression
  – description: TIFF file using lossy compression
  – class type: Policy
  – severity: 5

• Class IC041 – Forbidden TIA tags
  – description: Tiff with forbidden tags (for example: SubfileType, Thresholding, CellWidth, CellLength, FillOrder, MinSampleValue, MaxSampleValue, FreeOffsets, FreeByteCounts, T4Options, T6Options, TransferFunction, Predictor, WhitePoint, PrimaryChromaticities, ColorMap, HalftoneHints, TileWidth, TileLength, TileOffsets, TileByteCounts, SubIFDs, InkSet, InkNames, NumberOfInks, DotRange, TargetPrinter, ExtraSamples, SMinSampleValue, SMaxSampleValue, TransferRange, JPEGTables, JPEGProc, JPEGInterchangeFormat, JPEGInterchangeFormatLength, JPEGRestartInterval, JPEGLosslessPredictors, JPEGPointTransforms, JPEGQTables, JPEGDCTables, JPEGACTables, YCbCrCoefficients, YCbCrSubSampling, YCbCrPositioning, ReferenceBlackWhite, CFARrepeatPatternDim, CFAPattern, Interlace, CompressedBitsPerPixel, FocalPlaneXResolution, FocalPlaneYResolution, ImageSourceData)
  – class type: Policy
  – severity: 5

• Class IC042 – Mandatory TIA tags
  – description: Tiff without mandatory TIA tags (for example: NewSubfileType, ImageWidth, ImageLength, BitsPerSample, Compression, PhotometricInterpretation, StripOffsets, Orientation, SamplesPerPixel, RowsPerStrip, StripByteCounts, PlanarConfiguration) [ ]
- **class type**: Policy
- **severity**: 5

- **Class IC043 – Uncompressed Baseline IBM TIFF v6.0 RGB**
  - **description**: Image conformance TIFF Baseline 6.0 with little-endian byte order, RGB color without compression
  - **class type**: Policy
  - **severity**: 3

### A.3 Audio-video Media Type

- **Class AVC000 – Correct**
  - **description**: This class indicates the documents that do not have any conformance and/or policy issue.
  - **class type**: Conformance
  - **severity**: 0

- **Class AVC001 – The First Element must be the EBML Header.**
  - **description**: The first Element ID must equal 0x172351395 (EBML Header) [EBML/EBML-ELEM-START]
  - **class type**: Conformance
  - **severity**: 1

- **Class AVC002 – EBMLVersion must be greater than or equal to EBMLReadVersion**
  - **description**: EBMLReadVersion must be equal or less than the EBMLVersion. [EBML/EBML-VER-COH]
  - **class type**: Conformance
  - **severity**: 1

- **Class AVC003 – DocTypeVersion must be greater than or equal to DocTypeReadVersion**
  - **description**: DocTypeReadVersion must be equal or less than the DocTypeVersion. [EBML/EBML-DOCVER-COH]
  - **class type**: Conformance
  - **severity**: 1

- **Class AVC004 – All Elements MUST have valid parents**
  - **description**: Check that each EBML Element has a valid Parent Element. [EBML/EBML-ELEMENT-VALID-PARENT]
- **class type**: Conformance
- **severity**: 1

• **Class AV C005 – Elements follow maxOccurs**
  - **description**: Verify maxOccurs of EBML Elements [EBML/EBML-ELEMENT-NONMULTIPLES]
  - **class type**: Conformance
  - **severity**: 1

• **Class AV C006 – Elements follow minOccurs**
  - **description**: Verify minOccurs of EBML Elements [EBML/EBML-ELEMENT-CONTAINS-MANDATES]
  - **class type**: Conformance
  - **severity**: 1

• **Class AV C007 – EBMLMaxIDLength valid**
  - **description**: EBMLMaxIDLength must be in valid range. [EBML/EBML-VALID-MAXID]
  - **class type**: Conformance
  - **severity**: 1

• **Class AV C008 – EBMLMaxSizeLength valid**
  - **description**: EBMLMaxSizeLength must be in valid range. [EBML/EBML-VALID-MAXSIZE]
  - **class type**: Conformance
  - **severity**: 1

• **Class AV C011 – Header Elements in Element Data Size length range**
  - **description**: Element Data Size (descending from Root Element) lengths must be less than or equal to 4. [EBML/HEADER-ELEMENTS-WITHIN-MAXSIZELENGTH]
  - **class type**: Conformance
  - **severity**: 1

• **Class AV C012 – Elements in Element Data Size length range**
  - **description**: Element Data Size (descending from Root Element) lengths must be less than or equal to EBMLMaxSizeLength. [EBML/ELEMENTS-WITHIN-MAXSIZELENGTH]
  - **class type**: Conformance
  - **severity**: 1

• **Class AV C015 – Element Size 0x7F Reservation**
– **description**: Note that the shortest encoding form for 127 is 0x407f since 0x7f is reserved." If Element Size is set to 0x11111111 but element size is actually 127 bytes provide a warning." [EBML/EBML-ELEM-SIZE-7F]
– **class type**: Conformance
– **severity**: 1

• **Class AV C019 – Non-Ascii Data in String**
  – **description**: The string element is limited to certain byte ranges of ascii plus a trailing optional null byte. [EBML/EBML-NON-ASCII-IN-STRING]
  – **class type**: Conformance
  – **severity**: 1

• **Class AV C020 – Do the Matroska Seek Elements properly resolve**
  – **description**: Test offsets of Seek Elements to ensure they resolve properly. [EBML/MKV-SEEK-RESOLVE]
  – **class type**: Conformance
  – **severity**: 1

• **Class AV C021 – EBML CRC Element must be first**
  – **description**: The CRC Element if used must be the first Child Element of the Parent Element. [EBML/EBML-CRC-FIRST]
  – **class type**: Conformance
  – **severity**: 1

• **Class AV C022 – EBML CRC Element must contain a valid hash**
  – **description**: The stored CRC-32 value should verify. [EBML/EBML-CRC-VALID]
  – **class type**: Conformance
  – **severity**: 1

• **Class AV C024 – EBML Elements used correlate to DocVersion**
  – **description**: Elements defined with a specific minver should not be present in an EBML Document that uses an EBMLDocTypeVersion lower than that minver. [EBML/EBML-MINVER-COHERANT]
  – **class type**: Conformance
  – **severity**: 1

• **Class AV C025 – EBML Elements used correlate to DocVersion**
- **description**: Elements defined with a specific maxver should not be present in an EBML Document that uses an EBMLDocTypeVersion higher than that maxver. [EBML/EBML-MAXVER-COHERANT]
  - **class type**: Conformance
  - **severity**: 1

- **Class AV C028 – EBML Element adheres to size restrictions**
  - **description**: The length of the Element falls within the permitted range of the optional size declaration. [EBML/EBML-ELEMENT-IN-SIZE-RANGE]
  - **class type**: Conformance
  - **severity**: 1

- **Class AV C029 – EBML Element adheres to range restrictions**
  - **description**: The value of the Element falls within the permitted range. [EBML/EBML-ELEMENT-VALID-RANGE]
  - **class type**: Conformance
  - **severity**: 1

- **Class AV C030 – Matroska TrackType must be a valid value**
  - **description**: Only tracktype values of 1 2 3 16 17 18 32 are currently defined [Matroska/MKV-VALID-TRACKTYPE-VALUE]
  - **class type**: Conformance
  - **severity**: 1

- **Class AV C032 – Matroska Tags defined as numerical should be.**
  - **description**: Some tags are defined to be a number in a UTF-8 element test that the value is numeric. [Matroska/MKV-NUMERICAL-TAGS]
  - **class type**: Conformance
  - **severity**: 1

- **Class AV C033 – Missing header**
  - **description**: Version 2 and later files use a global header.” If version is 2 or more, there should be a global header in the container private data” [FFV1/OUTOFBAND-HEADER-MISSING]
  - **class type**: Conformance
  - **severity**: 1

- **Class AV C034 – version**
- **description**: version 0, 1 or 3. Maximum known version is 3 [FFV1/FFV1-HEADER-version]
  - **class type**: Conformance
  - **severity**: 1

- **Class AV C035 – version 2**
  - **description**: Version 2 was never enabled in the encoder thus version 2 files should not exist. Version 2 is forbidden - analysis stops [FFV1/FFV1-HEADER-version2]
  - **class type**: Conformance
  - **severity**: 1

- **Class AV C036 – micro_version 2**
  - **description**: For version 3, micro_version is 4 [FFV1/FFV1-HEADER-micro_version]
  - **class type**: Conformance
  - **severity**: 1

- **Class AV C037 – coder_type**
  - **description**: 0 (Golomb Rice), 1 (Range coder) [FFV1/FFV1-HEADER-coder_type]
  - **class type**: Conformance
  - **severity**: 1

- **Class AV C038 – state_transition_delta**
  - **description**: (To be defined), FFV1 [FFV1/FFV1-HEADER-state_transition_delta]
  - **class type**: Conformance
  - **severity**: 1

- **Class AV C039 – colorspace_type**
  - **description**: 0 (YCbCr), 1 (JPEG2000_RCT)" colorspace_type >1 is not supported" [FFV1/FFV1-HEADER-colorspace_type]
  - **class type**: Conformance
  - **severity**: 1

- **Class AV C040 – bits_per_raw_sample**
  - **description**: commonly 8, 9, 10, 12, 14, 16 [FFV1/FFV1-HEADER-bits_per_raw_sample]
  - **class type**: Conformance
  - **severity**: 1
• Class AVC041 – h_chroma_subsample
  – description: chroma subsampling factor can not be higher than slice width [FFV1/FFV1-HEADER-h_chroma_subsample-max]
  – class type: Conformance
  – severity: 1

• Class AVC042 – h_chroma_subsample
  – description: width divided by chroma subsampling factor is not an integer [FFV1/FFV1-HEADER-h_chroma_subsample-int]
  – class type: Conformance
  – severity: 1

• Class AVC043 – v_chroma_subsample
  – description: chroma subsampling factor can not be higher than slice height [FFV1/FFV1-HEADER-v_chroma_subsample-max]
  – class type: Conformance
  – severity: 1

• Class AVC044 – v_chroma_subsample
  – description: height divided by chroma subsampling factor is not an integer [FFV1/FFV1-HEADER-v_chroma_subsample-int]
  – class type: Conformance
  – severity: 1

• Class AVC045 – QuantizationTables
  – description: QuantizationTables incoherency [FFV1/FFV1-HEADER-QUANTIZATION_TABLES]
  – class type: Conformance
  – severity: 1

• Class AVC047 – ec
  – description: 0(32bit CRC on the global header), 1(32bit CRC per slice and the global header)* ec >1 is not supported* [FFV1/FFV1-HEADER-ec]
  – class type: Conformance
  – severity: 1

• Class AVC048 – intra
- **description**: intra 0(key and non key frames), 1(the video contains only key frames)" intra >1 is not supported" [FFV1/FFV1-HEADER-intra]
  - **class type**: Conformance
  - **severity**: 1

- **Class AV C049** – **crc_parity**
  - **description**: 32bit that are choosen so that the global header as a whole or slice as a whole has a crc" CRC is wrong" [FFV1/FFV1-HEADER-crc_parity]
  - **class type**: Conformance
  - **severity**: 1

- **Class AV C050** – **end of header**
  - **description**: Real header end is met before or after expected header end [FFV1/FFV1-HEADER-END]
  - **class type**: Conformance
  - **severity**: 1

- **Class AV C051** – **slice x / y / width / height**
  - **description**: Slices x/y and slices width/height are not coherant (areas are not sticked) [FFV1/FFV1-SLICE-slice_xywh]
  - **class type**: Conformance
  - **severity**: 1

- **Class AV C052** – **quant_table_index**
  - **description**: quant_table_index incoherency [FFV1/FFV1-SLICE-quant_table_index]
  - **class type**: Conformance
  - **severity**: 1

- **Class AV C053** – **picture_structure**
  - **description**: 0(unknown) 1(top field first) 2(bottom field first) 3(progressive)" picture_structure >3 is not supported" [FFV1/FFV1-SLICE-picture_structure]
  - **class type**: Conformance
  - **severity**: 1

- **Class AV C054** – **sar_den**
  - **description**: 0/0 when unknown " if num is not 0, den should be not 0" [FFV1/FFV1-SLICE-sar_den]
- **class type**: Conformance
  - **severity**: 1

- **Class AV C055 – slice_size**
  - **description**: slice_size is bigger than frame size [FFV1/FFV1-SLICE-slice_size]
  - **class type**: Conformance
  - **severity**: 1

- **Class AV C056 – error_status**
  - **description**: 0 (no error), 1 (slice contained a correctable error) [FFV1/FFV1-SLICE-crc_parity]
  - **class type**: Conformance
  - **severity**: 1

- **Class AV C057 – crc_parity**
  - **description**: 32bit that are chosen so that the global header as a whole or slice as a whole has a crc" CRC is wrong" [FFV1/FFV1-SLICE-crc_parity]
  - **class type**: Conformance
  - **severity**: 1

- **Class AV C058 – end of slice**
  - **description**: Real slice end is met before or after expected slice end [FFV1/FFV1-SLICE-END]
  - **class type**: Conformance
  - **severity**: 1

- **Class AV C059 – end of frame**
  - **description**: Real frame end is met before or after expected frame end [FFV1/FFV1-FRAME-END]
  - **class type**: Conformance
  - **severity**: 1

- **Class AV C061 – Matroska version 4 or greater?**
  - **description**: Is MKV at least version 4 [Matroska/MKV-V4+]
  - **class type**: Policy
  - **severity**: 1

- **Class AV C062 – SegmentUID is present?**
- **description**: A SegmentUID Element is stored. [Matroska/SEGMENTUID-PRESENT]
  - **class type**: Policy
  - **severity**: 1

- **Class AV C063 – SeekHead is present?**
  - **description**: A SeekHead Element is stored. [Matroska/SEEKHEAD-PRESENT]
  - **class type**: Policy
  - **severity**: 1

- **Class AV C064 – Interlaced video is clarified?**
  - **description**: Interlacement is set specifically even if unknown. [Matroska/INTERLACEMENT-CLARITY]
  - **class type**: Policy
  - **severity**: 1

- **Class AV C065 – Video Sample Range is clarified?**
  - **description**: Sample range is set specifically even if unknown. [Matroska/SAMPLE-RANGE-CLARITY]
  - **class type**: Policy
  - **severity**: 1

- **Class AV C066 – Video Colour Primary is clarified?**
  - **description**: Colour primary is set specifically even if unknown. [Matroska/COLOUR-PRIMARY-CLARITY]
  - **class type**: Policy
  - **severity**: 1

- **Class AV C067 – FFV1 is version 3 or greater**
  - **description**: FFV1 version 3 and greater is recommended for archival use since it adds self-descriptive and fixity features. [Matroska/FFV1-3+]
  - **class type**: Policy
  - **severity**: 1

- **Class AV C068 – If version 3, FFV1 is subversion 4 or greater**
  - **description**: FFV1 version 3 is only non-experimental in subversion 4 and higher. [Matroska/FFV1-3.4+]
  - **class type**: Policy
- severity: 1

- Class AV'C069 – **No junk data within Matroska**
  - description: All Master Elements only contain Elements. [Matroska/NO-JUNK-IN-MATROSKA]
  - class type: Policy
  - severity: 1

## B Created Ground-Truth

For each media type, we report the list of files used for test and which has been accessed through the PREFORMA Vault as well as the complete ground-truth for each class.

### B.1 Text Media Type

#### B.1.1 File List

- tr3.ar10.a1b.pdf
- Gam13_FINAL_Paper30_camera_ready.pdf
- action_resetform.ar10.pdf
- space.ar10.a1b.pdf
- transitions.ar10.a2u.pdf
- crypt.ar10.a2b.pdf
- goto3dview.ar10.a1b.pdf
- rolemap.ar10.a2a.pdf
- font_metric.ar10.a2u.pdf
- annotation_fileattachment.ar10.pdf
- attachment.ar10.pdf
- js.ar10.a2b.pdf
- attachment_pdf.ar10.a2b.pdf
- sound.ar10.a2u.pdf
- AJGI-PDF-017528_0104.pdf
- persian.mo03.ar10.a1a.pdf
- rm5.ar10.a2a.pdf
- attachment.ar10.a2a.pdf
- action_launch.ar10.pdf
- xmp_undefined.ar10.a1b.pdf
- base_unifont.ar10.a1b.pdf
- Microsoft Word - Riksarkivets beslut i äfrende 233-2015-4080.a1b.pdf
- 130325PREFORMAHangoutnotes250320131258.pdf
- attachment.ar10.a2b.pdf
- annotation_sound.ar10.a2b.pdf
- metadata_sync_keyw.MS010.internal.pdf
type.ar10.a2a.pdf
language_unknown.lo5.a1a.pdf
form_action_opendoc.ar10.pdf
transparency.ar10.a2b.pdf
permission.ar10.pdf
fileheader_15.ar10.a2b.pdf
resetform.ar10.ala.pdf
annotation_fileattachment_pdf.ar10.pdf
ccittg4.ar10.ala.pdf
action_goto3dview.ar10.pdf

B.1.2 Ground-truth

TC000 0 correct_tc054.win.1o5.ala.pdf 1
TC000 0 correct_win.odt.1o5.doc.mso3.aaxp.ala.pdf 1
TC000 0 correct_tc002_tc012_tc014_tc017_tc053.win.odt.1o5.doc.mso3.aaxp.a2u.pdf 1
TC000 0 correct_tc002_tc012_tc014_tc017.win.odt.1o5.doc.mso3.aaxp.a2a.pdf 1
TC000 0 correct_tc002_tc017.win.odt.1o5.doc.mso3.aaxp.a2a.pdf 1
TC000 0 correct_tc002_tc017.win.odt.1o5.doc.mso3.aaxp.a2u.pdf 1
TC000 0 correct_tc037_tc054.lin.1o5.ala.pdf 1
TC002 0 annotation_fileattachment_pdf.ar10.pdf 1
TC002 0 annotation_fileattachment.ar10.pdf 1
TC002 0 annotation_fileattachment_pdf.ar10.a2b.pdf 1
TC002 0 annotation_fileattachment.ar10.a2u.pdf 1
TC003 0 annotation_sound.ar10.a2a.pdf 1
TC003 0 annotation_sound.ar10.a2b.pdf 1
TC004 0 movie.ar10.ala.pdf 1
TC004 0 movie.ar10.a2b.pdf 1
TC004 0 action_movie.ar10.a2a.pdf 1
TC004 0 action_movie.ar10.pdf 1
TC005 0 screen.ar10.ala.pdf 1
TC005 0 screen.ar10.a2b.pdf 1
TC005 0 screen.ar10.a2u.pdf 1
TC006 0 action_goto3dview.ar10.pdf 1
TC006 0 goto3dview.ar10.a2a.pdf 1
TC006 0 goto3dview.ar10.ala.pdf 1
TC007 0 lzw.ar10.ala.pdf 1
TC007 0 lzw.ar10.a2b.pdf 1
TC007 0 lzw.ar10.a2u.pdf 1
TC008 0 encoding_crypt_edit_fxd.ar10.ala.pdf 1
TC008 0 crypt.ar10.ala.pdf 1
TC008 0 crypt.ar10.a2b.pdf 1
TC022 0 sound.ar10.a2b.pdf 1
TC022 0 sound.ar10.a2u.pdf 1
TC023 0 action_movie.ar10.a1a.pdf 1
TC023 0 action_movie.ar10.a2a.pdf 1
TC023 0 action_movie.ar10.pdf 1
TC024 0 action_hide.ar10.pdf 1
TC024 0 hide.a2b.pdf 1
TC024 0 hide.a2u.pdf 1
TC024 0 hide.a1a.pdf 1
TC025 0 action_resetform.ar10.pdf 1
TC025 0 resetform.ar10.a1a.pdf 1
TC025 0 resetform.ar10.a2b.pdf 1
TC025 0 resetform.ar10.a2u.pdf 1
TC026 0 action_importdata.ar10.pdf 1
TC026 0 importdata.ar10.a1b.pdf 1
TC026 0 importdata.ar10.a2b.pdf 1
TC026 0 importdata.ar10.a2u.pdf 1
TC027 0 action_javascript.ar10.pdf 1
TC027 0 js.ar10.a1a.pdf 1
TC027 0 js.ar10.a2b.pdf 1
TC027 0 js.ar10.a2u.pdf 1
TC028 0 setstate.ar10.a1b.pdf 1
TC028 0 setstate.ar10.a2a.pdf 1
TC028 0 setstate.ar10.a2u.pdf 1
TC029 0 noop.ar10.a1b.pdf 1
TC029 0 noop.ar10.a2a.pdf 1
TC029 0 noop.ar10.a2u.pdf 1
TC030 0 setocgstate.ar10.a1b.pdf 1
TC030 0 setocgstate.ar10.a2a.pdf 1
TC030 0 setocgstate.ar10.a2b.pdf 1
TC030 0 setocgstate.ar10.a2u.pdf 1
TC031 0 action_rendition.ar10.a2b.pdf 1
TC031 0 action_rendition.ar10.a2u.pdf 1
TC032 0 action_trans.ar10.pdf 1
TC032 0 action_trans.ar10.a1b.pdf 1
TC032 0 action_trans.ar10.a2a.pdf 1
TC033 0 action_goto3dview.ar10.pdf 1
TC033 0 goto3dview.ar10.a2a.pdf 1
TC033 0 goto3dview.ar10.a1b.pdf 1
TC034 0 named.ar10.a1a.pdf 1
TC034 0 named.ar10.a2b.pdf 1
TC035 0 encryption.ar10.a1b.pdf 1
TC035 0 encryption.ar10.a2a.pdf 1
TC043 0 wingdings.pdf.ar10.a2u.pdf 1
TC044 0 rolemap.ar10.a2a.pdf 1
TC044 0 rm5.ar10.a2a.pdf 1
TC044 0 rm3.ar10.a1a.pdf 1
TC044 0 rm6.ar10.a1a.pdf 1
TC045 0 130325PREFORMAHangoutnotes250320131258.pdf 1
TC045 0 141191_1875.pdf 1
TC045 0 20140227_3_1.pdf 1
TC045 0 20140401_117_1.pdf 1
TC045 0 20141201_6_106_1.pdf 1
TC045 0 285666_001_b12034058_0001.pdf 1
TC045 0 AJGI-PDF-017401_0103.pdf 1
TC045 0 AJGI-PDF-017401_0104.pdf 1
TC045 0 AJGI-PDF-017401_0105.pdf 1
TC045 0 AJGI-PDF-017401_0106.pdf 1
TC045 0 AJGI-PDF-017693_0100.pdf 1
TC045 0 AJGI-PDF-017693_0104.pdf 1
TC045 0 AJGI-PDF-017693_0106.pdf 1
TC045 0 AktÄr eller rum.pdf 1
TC045 0 Evaluation report.pdf 1
TC045 0 KarlPetterssonPMDigitaltBevarandeA.pdf 1
TC045 0 Microsoft Word - Riksarkivets beslut i Ådrende 233-2015-4080.pdf 1
TC045 0 TC_06435.pdf 1
TC045 0 TC_06437.pdf 1
TC045 0 TC_06439.pdf 1
TC045 0 TC_06441.pdf 1
TC045 0 TC_06443.pdf 1
TC045 0 Word_PDFa_Distiller_teckensnitt_omslutalldt_(tsnitt omslut).pdf 1
TC045 0 Word_teckensnitt-omsluten_intePDFa_Distiller_(tsnitt !omslut).pdf 1
TC045 0 __TMP__.pdf 1
TC045 0 b2.ar10.ocr-exact.a1b.pdf 1
TC045 0 b2026033x.pdf 1
TC045 0 ccittg4.ar10.a1b.pdf 1
TC045 0 ccittg4.ar10.pdf 1
TC045 0 dokument.lo4.odg.jpg.ar10.a1b.pdf 1
TC045 0 dom20170918.pdf 1
TC045 0 horisont_2009_06.pdf 1
TC045 0 image_per_page.ar10.a2u.pdf 1
TC045 0 permission.ar10.pdf 1
TC045 0 presentation.odp.lo5.pdf 1
TC045 0 protected.ar10.a1a.pdf 1
TC045 0 protected.ar10.a2b.pdf 1
TC050 0 annotation_fileattachment_pdf.ar10.a2u.pdf 1
TC050 0 annotation_fileattachment_pdfa.ar10.a2a.pdf 1
TC050 0 annotation_sound.ar10.a2a.pdf 1
TC050 0 annotation_sound.ar10.a2b.pdf 1
TC050 0 arabic.lo5.ar10.a2a.pdf 1
TC050 0 artifact.ar10.a2a.pdf 1
TC050 0 attachment_odt.ar10.a2a.pdf 1
TC050 0 attachment_pdf.ar10.a2a.pdf 1
TC050 0 attachment_pdf.ar10.a2b.pdf 1
TC050 0 attachment_pdf.ar10.a2u.pdf 1
TC050 0 attachment_pdfa.ar10.a2a.pdf 1
TC050 0 ccittg4.ar10.pdf 1
TC050 0 character_identification_of_font_subsets.mso03.ar10.a2b.pdf 1
TC050 0 character_identification_of_font_subsets.mso03.ar10.a2u.pdf 1
TC050 0 correct.odt.lo5.doc.mso03.aaxp.pdf 1
TC050 0 correct_tc037_tc054.win.odt.lo5.doc.mso03.aaxp.a2u.pdf 1
TC050 0 crypt.ar10.a2u.pdf 1
TC050 0 digital_signatur.MSO03.AR10.1.7.a2u.pdf 1
TC050 0 dom20170918.pdf 1
TC050 0 encoding_crypt_edit_fxd.ar10.a1a.pdf 1
TC050 0 file_attachment_xml.ar10.a2b.pdf 1
TC050 0 file_attachment_xml.ar10.a2u.pdf 1
TC050 0 fileheader_41.ar10.a2a.pdf 1
TC050 0 form_action_opendoc.ar10.a2a.pdf 1
TC050 0 fwidth_charset.ar10.pdf 1
TC050 0 hebrew.mo03.ar10.a2a.pdf 1
TC050 0 hebrew.mo03.ar10.a2b.pdf 1
TC050 0 hebrew.mo03.ar10.a2u.pdf 1
TC050 0 hindi.mo03.ar10.a2a.pdf 1
TC050 0 hindi.mo03.ar10.a2b.pdf 1
TC050 0 hindi.mo03.ar10.a2u.pdf 1
TC050 0 image_per_page.ar10.a2a.pdf 1
TC050 0 image_per_page.ar10.a2b.pdf 1
TC050 0 image_per_page.ar10.a2u.pdf 1
TC050 0 japanese_v1.mo03.ar10.a2u.pdf 1
TC050 0 javascript.ar10.a2a.pdf 1
TC050 0 javascript.ar10.a2b.pdf 1
TC050 0 keyword_spacing_stream.ar10.a2b.pdf 1
TC050 0 lager.L04.0.ACR10.1.7.a2u.pdf 1
TC050 0 metadata.MSO10.internal.AR10.1.7.a2u.pdf 1
TC050 0 metadata.MSO10.internal.AR10.1.7.pdf 1
TC050 0 metadata_sync_crea.MSO10.internal.AR10.1.7.a2u.pdf 1
TC050 0 notdef.ar10.a1a.pdf 1
TC056 0 tr3.ar10.a1b.pdf 1
TC056 0 tr3.ar10.a2a.pdf 1
TC058 0 character_identification_of_font_subsets.mso03.ar10.a1a.pdf 1
TC059 0 Moodlemoot_2014_fullpaper_ITA_CAB_UNIPD.pdf 1
TC059 0 Progetto_Tecnico_Municipio_RM5.pdf 1
TC059 0 Valori_Storici_V_Municipio.pdf 1
TC059 0 __TEMPLATE__.lo5.a1a.pdf 1
TC059 0 action_goto3dview.ar10.a1a.pdf 1
TC059 0 action_goto3dview.ar10.pdf 1
TC059 0 action_hide.ar10.pdf 1
TC059 0 action_hide.lo5.a1a.pdf 1
TC059 0 action_importdata.ar10.a1a.pdf 1
TC059 0 action_importdata.lo5.a1a.pdf 1
TC059 0 action_importdata.ar10.pdf 1
TC059 0 action_importdata.ar10.pdf 1
TC059 0 action_javascript.ar10.a1a.pdf 1
TC059 0 action_javascript.ar10.pdf 1
TC059 0 action_launch.ar10.pdf 1
TC059 0 action_launch.lo5.a1a.pdf 1
TC059 0 action_movie.ar10.pdf 1
TC059 0 action_named.lo5.a1a.pdf 1
TC059 0 action_named_nextpage.ar10.pdf 1
TC059 0 action_rendition.lo5.a1a.pdf 1
TC059 0 action_resetform.ar10.pdf 1
TC059 0 action_setocgstate.ar10.pdf 1
TC059 0 action_setocgstate.lo5.a1a.pdf 1
TC059 0 action_sound.ar10.pdf 1
TC059 0 action_trans.ar10.pdf 1
TC059 0 alternativa_beskrivningar.LO4.0.internal.pdf 1
TC059 0 alternativa_beskrivningar.MSO03.AR10.1.7.pdf 1
TC059 0 alternativa_beskrivningar.MSO10.internal.pdf 1
TC059 0 annotation_file_attachment_xml.ar10.pdf 1
TC059 0 annotation_fileattachment.ar10.pdf 1
TC059 0 annotation_fileattachment_pdf.ar10.a1a.pdf 1
TC059 0 annotation_fileattachment_pdf.ar10.pdf 1
TC059 0 annotation_fileattachment_pdf.ar10.pdf 1
TC059 0 annotation_fileattachment_pdf.ar10.pdf 1
TC059 0 annotation_sound.lo5.a1a.pdf 1
TC059 0 arabic.mo03.ar10.a1a.pdf 1
TC059 0 arabic.mo03.ar10.a1a.pdf 1
TC059 0 attachment.ar10.a1a.pdf 1
TC059 0 attachment.ar10.pdf 1
TC059 0 attachment_odt.lo5.a1a.pdf 1
TC059 0 attachment_pdf.lo5.a1a.pdf 1
TC059 0 attachment_pdfa.ar10.a1a.pdf 1
TC059 0 jpx.lo5.a1a.pdf 1
TC059 0 js.ar10.a1a.pdf 1
TC059 0 keyword_spacing_xref.lo5.a1a.pdf 1
TC059 0 komprimering.MSO03.AR10.1.7.pdf 1
TC059 0 korean_hangul_v1.lo5.a1a.pdf 1
TC059 0 korean_hangul_v2.mo03.ar10.a1a.pdf 1
TC059 0 lager.MSO03.AR10.1.7.pdf 1
TC059 0 language_unknown.lo5.a1a.pdf 1
TC059 0 metadata.LO4.0.internal.pdf 1
TC059 0 metadata.MO03.AR10.1.7.pdf 1
TC059 0 metadata.MSO10.internal.pdf 1
TC059 0 metadata_sync_auth.MO03.AR10.1.7.pdf 1
TC059 0 metadata_sync_keyw.MSO10.internal.pdf 1
TC059 0 movie.ar10.a1a.pdf 1
TC059 0 named.ar10.a1a.pdf 1
TC059 0 no_metadata.lo5.a1a.pdf 1
TC059 0 notdef.lo5.a1a.pdf 1
TC059 0 permission.lo5.a1a.pdf 1
TC059 0 persian.mo03.ar10.a1a.pdf 1
TC059 0 rapport_kompilat.mso03-ar10.doc.a1a.pdf 1
TC059 0 resetform.ar10.a1a.pdf 1
TC059 0 rm3.ar10.a1a.pdf 1
TC059 0 rm6.ar10.a1a.pdf 1
TC059 0 screen.ar10.a1a.pdf 1
TC059 0 servizi_centralizzati_2015_Zane_short_conBettella_DEF.pdf 1
TC059 0 skylt.lo4.4.docx.a1a.pdf 1
TC059 0 slutrapport.MSO2010.docx.a1a.pdf 1
TC059 0 sound.ar10.a1a.pdf 1
TC059 0 swedish.mo3.ar10.a1a.pdf 1
TC059 0 test-osynlig-underskrift.lo4.4.odt.a2a.pdf 1
TC059 0 test-synlig-underskrift.lo4.4.odt.a1a.pdf 1
TC059 0 text-whiteout.lo5.odt.a1a.pdf 1
TC059 0 text_rendering_mode3.lo5.a1a.pdf 1
TC059 0 text_rendering_mode3_not-embedded.ar10.a1a.pdf 1
TC059 0 thoms_1931_ireland_pg0008.pdf 1
TC059 0 transitions.ar10.a1a.pdf 1
TC059 0 transitions.ar10.pdf 1
TC059 0 transparency.ar10.a1b.pdf 1
TC059 0 transparency.ar10.pdf 1
TC059 0 transparency.lo5.a1a.pdf 1
TC059 0 utÅ¾kad text.LO4.0.internal.AR10.1.7.pdf 1
TC059 0 utÅ¾kad text.MSO03.AR10.1.7.AR10.1.7.pdf 1
| TC065 0 annotation_fileattachment_pdf.ar10.pdf 1 |
| TC065 0 annotation_fileattachment_pdfa.ar10.a1a.pdf 1 |
| TC065 0 annotation_sound.lo5.ala.pdf 1 |
| TC065 0 arabic.mo03.ar10.ala.pdf 1 |
| TC065 0 armenian.lo5.ala.pdf 1 |
| TC065 0 attachment.ar10.ala.pdf 1 |
| TC065 0 attachment.ar10.a2a.pdf 1 |
| TC065 0 attachment.ar10.pdf 1 |
| TC065 0 attachment_odt.lo5.ala.pdf 1 |
| TC065 0 attachment_pdf.lo5.ala.pdf 1 |
| TC065 0 attachment_pdfa.ar10.ala.pdf 1 |
| TC065 0 attachment_pdfa.ar10.pdf 1 |
| TC065 0 attachment_pdfa.lo5.ala.pdf 1 |
| TC065 0 attachment.ar10.a1a.pdf 1 |
| TC065 0 attachment.ar10.a1a.pdf 1 |
| TC065 0 attachment.ar10.pdf 1 |
| TC065 0 attachment_odt.lo5.ala.pdf 1 |
| TC065 0 attachment_pdf.lo5.ala.pdf 1 |
| TC065 0 attachment_pdfa.ar10.ala.pdf 1 |
| TC065 0 attachment_pdfa.ar10.pdf 1 |
| TC065 0 attachment_pdfa.lo5.ala.pdf 1 |
| TC065 0 base_dejavusans.ar10.ala.pdf 1 |
| TC065 0 base_times.ar10.ala.pdf 1 |
| TC065 0 base_unifont.ar10.ala.pdf 1 |
| TC065 0 character_identification_of_font_subsets.mso03.ar10.ala.pdf 1 |
| TC065 0 correct_win.odt.lo5.doc.mso3.aaxp.pdf 1 |
| TC065 0 correct_tc002_tc012_tc014_tc017_tc053.win.odt.lo5.doc.mso3.aaxp.pdf 1 |
| TC065 0 correct_tc037_tc054.lin.lo5.ala.pdf 1 |
| TC065 0 correct_tc054.win.lo5.ala.pdf 1 |
| TC065 0 crypt.ar10.ala.pdf 1 |
| TC065 0 digital_signatur.LO4.0.internal.pdf 1 |
| TC065 0 digital_signatur.MSO03.AR10.1.7.certifiera-osynlig_underskrift2.pdf 1 |
| TC065 0 effekt.LO4.0.internal.pdf 1 |
| TC065 0 effekt.MSO10.internal.pdf 1 |
| TC065 0 encoding_crypt.ar10.ala.pdf 1 |
| TC065 0 encoding_crypt.lo5.ala.pdf 1 |
| TC065 0 encoding_lzw.lo5.ala.pdf 1 |
| TC065 0 eof.lo5.ala.pdf 1 |
| TC065 0 faktisk_text.LO4.0.internal.AR10.1.7.pdf 1 |
| TC065 0 faktisk_text.MSO10.internal.pdf 1 |
| TC065 0 file_attachment_xml.ar10.ala.pdf 1 |
| TC065 0 fileheader_00.ar10.a2a.pdf 1 |
| TC065 0 fileheader_12.ar10.pdf 1 |
| TC065 0 fileheader_17.ar10.ala.pdf 1 |
| TC065 0 fileheader_20.ar10.pdf 1 |
| TC065 0 fileheader_20.lo5.ala.pdf 1 |
| TC065 0 font_metric.lo5.ala.pdf 1 |
TC065 0 form_action_opendoc.a1a.pdf 1
TC065 0 form_action_opendoc.ar10.a1a.pdf 1
TC065 0 form_action_opendoc.ar10.pdf 1
TC065 0 goto3dview.ar10.a2a.pdf 1
TC065 0 greek.mo03.ar10.a1a.pdf 1
TC065 0 hebrew.lo5.a1a.pdf 1
TC065 0 hide.a1a.pdf 1
TC065 0 hierarkisk struktur.LO4.0.internal.pdf 1
TC065 0 hierarkisk struktur.MSO03.AR10.1.7.pdf 1
TC065 0 hierarkisk struktur.MSO10.internal.pdf 1
TC065 0 japanese_jugemu.mo03.ar10.a1a.pdf 1
TC065 0 japanese_v1.lo5.a1a.pdf 1
TC065 0 japanese_v2.mo03.ar10.a1a.pdf 1
TC065 0 javascript.ar10.a1a.pdf 1
TC065 0 javascript.ar10.pdf 1
TC065 0 jpx.lo5.a1a.pdf 1
TC065 0 js.ar10.a1a.pdf 1
TC065 0 js.lo5.a2a.pdf 1
TC065 0 keyword_spacing_xref.lo5.a1a.pdf 1
TC065 0 komprimering.MSO03.AR10.1.7.pdf 1
TC065 0 korean_hangul_v1.lo5.a1a.pdf 1
TC065 0 korean_hangul_v2.mo03.ar10.a1a.pdf 1
TC065 0 lager.MSO03.AR10.1.7.pdf 1
TC065 0 language_unknown.lo5.a1a.pdf 1
TC065 0 launch.ar10.a2a.pdf 1
TC065 0 metadata.LO4.0.internal.pdf 1
TC065 0 metadata.MO03.AR10.1.7.pdf 1
TC065 0 metadata.MSO10.internal.pdf 1
TC065 0 metadata.MSO10.1.7.pdf 1
TC065 0 metadata.Sync.Auth.MSO03.AR10.1.7.pdf 1
TC065 0 metadata.Sync.Keyw.MSO10.internal.pdf 1
TC065 0 movie.ar10.a1a.pdf 1
TC065 0 movie.ar10.a2b.pdf 1
TC065 0 named.ar10.a1a.pdf 1
TC065 0 no_metadata.lo5.a1a.pdf 1
TC065 0 noop.ar10.a2a.pdf 1
TC065 0 notdef.lo5.a1a.pdf 1
TC065 0 permission.lo5.a1a.pdf 1
TC065 0 persian.mo03.ar10.a1a.pdf 1
TC065 0 rapport_kompilat.mso03-ar10.doc.a1a.pdf 1
TC065 0 resetform.ar10.a1a.pdf 1
TC065 0 rm3.ar10.a1a.pdf 1
TC065 0 rm5.ar10.a2a.pdf 1
TC068 0 ccittg4.ar10.a2a.pdf 1
TC068 0 jpx.ar10.a2u.pdf 1
TC069 0 20140401_117_1.pdf 1
TC070 0 ccittg4.ar10.a1b.pdf 1
TC070 0 ccittg4.ar10.a2a.pdf 1
TC070 0 ccittg4.ar10.pdf 1
TC071 0 AJGI-PDF-017401_0106.pdf 1
TC071 0 AJGI-PDF-017401_0103.pdf 1
TC071 0 komprimering.MSD03.AR10.1.7.pdf 1
TC071 0 encoding_lzw.ar10.a2b.pdf 1
TC071 0 encoding_lzw.ar10.a2u.pdf 1
TC071 0 encoding_lzw.lo5.a1a.pdf 1
TC073 0 no_metadata.ar10.a2b.pdf 1
TC073 0 no_metadata.ar10.a2u.pdf 1
TC073 0 no_metadata.lo5.a1a.pdf 1
TC073 0 AJGI-PDF-017401_0105.pdf 1
TC074 0 font_copyright.ar10.a1b.pdf 1
TC074 0 font_copyright.ar10.a2a.pdf 1
TC074 0 font_copyright.ar10.a2u.pdf 1
TC075 0 Microsoft Word - Riksarkivets beslut i Åfrende 233-2015-4080.a1b.pdf 1
TC075 0 horisont_2009_06.pdf 1
TC075 0 action_goto3dview.ar10.pdf 1
TC077 0 dokument.lo4.odg.jpg.arp10.a1b.pdf 1
TC077 0 b2026033x.pdf 1
TC078 0 no_metadata.ar10.a2b.pdf 1
TC078 0 no_metadata.ar10.a2u.pdf 1
TC078 0 no_metadata.lo5.a1a.pdf 1
TC078 0 AJGI-PDF-017528_0104.pdf 1
TC079 0 image_per_page.ar10.a1a.pdf 1
TC079 0 image_per_page.ar10.a2a.pdf 1
TC079 0 image_per_page.ar10.a2b.pdf 1
TC079 0 image_per_page.ar10.a2u.pdf 1
TC079 0 horisont_2009_06.pdf 1
TC079 0 TC_06443.pdf 1
TC080 0 ccittg4.ar10.a1b.pdf 1
TC080 0 ccittg4.ar10.a2a.pdf 1
TC080 0 ccittg4.ar10.pdf 1
TC081 0 b2.ar10.ocr-exact.a1b.pdf 1
TC082 0 AJGI-PDF-017528_0104.pdf 1
TC082 0 AJGI-PDF-017401_0103.pdf 1
TC082 0 AJGI-PDF-017528_0106.pdf 1
TC083 0 dokumentsprÅék.L04.0.internal.pdf 1
B.2 Image Media Type

B.2.1 File List

IMG_KOBigTiff.tif
Image50307yc.tif
pc260001.tif
Image30307j.tif
Image10307.tif
Image50307yc.tif
Image30304.tif
BigTIFFMotorola.tif
IMG_KO_overlap2.tif
DSC03635_ko_cardinality 10.tif
IMG_KO_date4.tif
Image40306.tif
BilevelCC.tif
BigTIFFSubIFD4.tif
IMG_OK.tif
DSCF10002P.tif
Image20306.tif
00160792_00000694.tif
SANY0038_ko_spp4.tif
IndexedEP.tif
Image40308.tif
Image20309.tif
IndexedEP.tif
Planar tile2.tif
SANY0038.tif
DSC03635_ko_cardinality 5.tif
Image30307j.tif
Planar tileB.tif
DSCF10002C.tif
DSC03635_ko_cardinality 5.tif
Image30306.tif
Image50307t2.tif
DSC03635_ko_cardinality 8.tif
Image60307_1.tif
Image20307.tif
Image60304g.tif
BigTIFFMotorolaLongStrips.tif
IMG_KO no ifds2.tif
BigTIFFSubIFD4.tif
00160792_00000692.tif
neuschwanstein kostrips.tif
SANY0038 ko x.tif
IMG_KO date5.tif
neuschwanstein.tif
neuschwanstein kostrips.tif
Image40307.tif
IMG_KO_pointzero.tif
Image40307.tif
DSCF10002CI.tif
Image50307yc.tif
IMG_OK.tif
SANY0038 ko photo 4.tif
DSCF10002b.tif
montserrat.tif
BigTIFFMotorola.tif
DSC03635_ko_tagtype 7.tif
neuschwanstein kostrips 2.tif
Image40304-6b.tif
00160792_00000695.tif
DSC03635_ko_tagtype 6.tif
Image20309.tif
Planar tile2.tif
IMG_KO date3.tif
IMG_KO_pointzero.tif
DSCF10002 ko strips 2.tif
Image10308.tif
Image10309.tif
Image40304-6.tif
BigTIFFSubIFD8.tif
DSC03635_ko_cardinality 10.tif
BilevelCC.tif
IndexedEP.tif
montserrat2.tif

B.2.2 Ground-truth

IC000 0 IMG_OK_channels.tif 1
IC000 0 DSC03635.tif 1
IC000 0 DSCF10002.tif 1
IC000 0 DSCF10002b.tif 1
IC000 0 montserrat2.tif 1
IC001 0 BigTIFFMotorola.tif 1
IC001 0 BigTIFFMotorolaLongStrips.tif 1
IC001 0 BigTIFFSubIFD4.tif 1
IC001 0 BigTIFFSubIFD8.tif 1
IC001 0 IMG_KOBigTiff.tif 1
IC001 0 Image30309.tif 1
IC001 0 Image30309b.tif 1
IC001 0 DSCF10002b_ko_magic.tif 1
IC002 0 Planar tile2.tif 1
IC002 0 00160792_00000695.tif 1
IC002 0 DSC03635_ko_cardinality 7.tif 1
IC002 0 DSCF10002_ko_type 2.tif 1
IC002 0 neuschwanstein kosrips.tif 1
IC002 0 DSC03635_ko_cardinality 9.tif 1
IC002 0 SANY0038 ko spp4.tif 1
IC002 0 DSC03635_ko_tagtype 7.tif 1
IC002 0 Image30304.tif 1
IC002 0 neuschwanstein kosrips 2.tif 1
IC002 0 DSCF10002 ko spp.tif 1
IC002 0 DSC03635_ko_cardinality 6.tif 1
IC002 0 IMG_KO date4.tif 1
IC002 0 DSC03635_ko_tagtype 10.tif 1
IC002 0 00160792_00000696.tif 1
<table>
<thead>
<tr>
<th>IC022</th>
<th>Image50307t2.tif 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>IC022</td>
<td>Planar_tile2.tif 1</td>
</tr>
<tr>
<td>IC023</td>
<td>Image50307t2.tif 1</td>
</tr>
<tr>
<td>IC024</td>
<td>Image30307t2.tif 1</td>
</tr>
<tr>
<td>IC025</td>
<td>Planar_tileB.tif 1</td>
</tr>
<tr>
<td>IC026</td>
<td>BilevelCC.tif 1</td>
</tr>
<tr>
<td>IC027</td>
<td>Image60304g.tif 1</td>
</tr>
<tr>
<td>IC028</td>
<td>DSCF10002P.tif 1</td>
</tr>
<tr>
<td>IC029</td>
<td>DSCF10002T.tif 1</td>
</tr>
<tr>
<td>IC030</td>
<td>DSCF10002C.tif 1</td>
</tr>
<tr>
<td>IC031</td>
<td>DSCF10002Y.tif 1</td>
</tr>
<tr>
<td>IC032</td>
<td>DSCF10002CI.tif 1</td>
</tr>
<tr>
<td>IC033</td>
<td>DSCF10002R.tif 1</td>
</tr>
<tr>
<td>IC033</td>
<td>montserrat.tif 1</td>
</tr>
<tr>
<td>IC034</td>
<td>00160792_00000692.tif 1</td>
</tr>
<tr>
<td>IC034</td>
<td>00160792_00000694.tif 1</td>
</tr>
<tr>
<td>IC034</td>
<td>00160792_00000695.tif 1</td>
</tr>
<tr>
<td>IC034</td>
<td>00160792_00000696.tif 1</td>
</tr>
<tr>
<td>IC034</td>
<td>Image30306.tif 1</td>
</tr>
<tr>
<td>IC034</td>
<td>IMG_KD_pointzero.tif 1</td>
</tr>
<tr>
<td>IC034</td>
<td>IMG_KD_date4.tif 1</td>
</tr>
<tr>
<td>IC034</td>
<td>DSC03635_ko_cardinality 10.tif 1</td>
</tr>
<tr>
<td>IC035</td>
<td>00160792_00000696.tif 1</td>
</tr>
<tr>
<td>IC035</td>
<td>IMG_KD_date2.tif 1</td>
</tr>
<tr>
<td>IC035</td>
<td>IMG_KD_date3.tif 1</td>
</tr>
<tr>
<td>IC035</td>
<td>IMG_KD_date4.tif 1</td>
</tr>
<tr>
<td>IC035</td>
<td>IMG_KD_date5.tif 1</td>
</tr>
<tr>
<td>IC035</td>
<td>00160792_00000695.tif 1</td>
</tr>
<tr>
<td>IC035</td>
<td>00160792_00000694.tif 1</td>
</tr>
<tr>
<td>IC035</td>
<td>00160792_00000692.tif 1</td>
</tr>
<tr>
<td>IC035</td>
<td>DSC03635_ko_cardinality 10.tif 1</td>
</tr>
<tr>
<td>IC036</td>
<td>montserrat.tif 1</td>
</tr>
<tr>
<td>IC037</td>
<td>IMG_OK.tif 1</td>
</tr>
<tr>
<td>IC037</td>
<td>KLEURTEST__ALT_LIDE120.tif 1</td>
</tr>
<tr>
<td>IC037</td>
<td>neuschwanstein.tif 1</td>
</tr>
<tr>
<td>IC037</td>
<td>pc260001.tif 1</td>
</tr>
<tr>
<td>IC037</td>
<td>sample.tif 1</td>
</tr>
<tr>
<td>IC037</td>
<td>SANY0038.tif 1</td>
</tr>
<tr>
<td>IC037</td>
<td>IndexedEP.tif 1</td>
</tr>
<tr>
<td>IC038</td>
<td>IndexedEP.tif 1</td>
</tr>
<tr>
<td>IC039</td>
<td>pc260001.tif 1</td>
</tr>
<tr>
<td>IC039</td>
<td>hgfhdgdf.tif 1</td>
</tr>
</tbody>
</table>
B.3 Audio-video Media Type

B.3.1 File List

MKV-V4.xml
EBML-VALID-MAXSIZE.mkv
FFV1-SLICE-slice_xywh_w_v3.mkv
FFV1-HEADER-version-EXPERIMENTAL.mkv
MKV-SAMPLE-RANGE-CLARITY.mkv
MKV-INTERLACEMENT-CLARITY.mkv
MKV-COLOUR-PRIMARY-CLARITY.xml
EBML-ELEMENT-NONMULTIPLES.mkv
FFV1-SLICE-error_status_v3.mkv
FFV1-HEADER-state_transition_delta_MS_v0.mkv
FFV1-HEADER-num_h_slices_MS_v3.mkv
EBML-ELEM-START.xml
FFV1-HEADER-bits_per_raw_sample_MS_v1.mkv
MKV-SEGMENTUID-PRESENT.xml
FFV1-SLICE-SliceContent_v3_GR.mkv
FFV1-SLICE-slice_xywh_h_v3.mkv
Correct_MS_v3_RC.mkv
EBML-ELEMENT-IN-SIZE-RANGE.mkv
MKV-SEEK-RESOLVE.mkv
FFV1-SLICE-SliceContent_v3_RC.mkv
FFV1-HEADER-state_transition_delta_MS_v3.mkv
Correct_MS_v0_GR.mkv
EBML-ELEMENT-CONTAINS-MANDATES.mkv
EBML-ELEM-START.mkv
EBML-VER-COH.mkv
FFV1-HEADER-configuration_record_crc_parity_v3.mkv
FFV1-HEADER-colorspace_type_MS_v0.mkv
FFV1-HEADER-version-OUTOFBAND_MS_v3.mkv
FFV1-HEADER-version-OUTOFBAND_MS_v0.mkv
B.3.2 Ground-truth

AVC000 0 Correct_MS_v0_GR.mkv 1
AVC000 0 Correct_MS_v0_RC.mkv 1
AVC000 0 Correct_MS_v3_GR.mkv 1
AVC000 0 Correct_MS_v3_RC.mkv 1
AVC001 0 EBML-ELEM-START.xml 1
AVC001 0 EBML-ELEM-START.mkv 1
AVC002 0 EBML-VER-COH.mkv 1
AVC003 0 EBML-DOCVREF-COH.mkv 1
AVC004 0 MKV-ELEMENT-VALID-PARENT.mkv 1
AVC005 0 EBML-ELEMENT-NONMULTIPLES.mkv 1
AVC006 0 EBML-ELEMENT-CONTAINS-MANDATES.mkv 1
AVC007 0 EBML-VALID-MAXID.mkv 1
AVC008 0 EBML-VALID-MAXSIZE.mkv 1
AVC011 0 EBML-HEADER-ELEMENTS-WITHIN-MAXSIZELENGTH.mkv 1
AVC012 0 EBML-ELEMENTS-WITHIN-MAXSIZELENGTH.mkv 1
AVC015 0 EBML-ELEM-UNKNOWN-SIZE.mkv 1
AVC019 0 EBML-ASCII-ONLY-IN-STRING.mkv 1
AVC020 0 MKV-SEEK-RESOLVE.mkv 1
AVC021 0 EBML-CRC-FIRST.mkv 1
AVC022 0 EBML-CRC-VALID.mkv 1
AVC024 0 EBML-MINVER-COHERANT.mkv 1
AVC025 0 EBML-MAXVER-COHERANT.mkv 1
AVC028 0 EBML-ELEMENT-IN-SIZE-RANGE.mkv 1
AVC029 0 EBML-ELEMENT-VALID-RANGE.mkv 1
AVC030 0 MKV-VALID-TRACKTYPE-VALUE.mkv 1
AVC032 0 MKV-NUMERICAL-TAG.mkv 1
AVC033 0 FFV1-HEADER-version-OUTOFBAND_MS_v0.mkv 1
AVC033 0 FFV1-HEADER-version-OUTOFBAND_MS_v3.mkv 1
AVC034 0 FFV1-HEADER-version-LATERVERSION.mkv 1
AVC035 0 FFV1-HEADER-version-EXPERIMENTAL.mkv 1
AVC036 0 FFV1-HEADER-micro_version_MS_v3.mkv 1
AVC037 0 FFV1-HEADER-coder_type_MS_v0.mkv 1
AVC037 0 FFV1-HEADER-coder_type_MS_v3.mkv 1
AVC038 0 FFV1-HEADER-state_transition_delta_MS_v0.mkv 1
AVC038 0 FFV1-HEADER-state_transition_delta_MS_v3.mkv 1
AVC039 0 FFV1-HEADER-colorspace_type_MS_v0.mkv 1
AVC039 0 FFV1-HEADER-colorspace_type_MS_v3.mkv 1
AVC040 0 FFV1-HEADER-bits_per_raw_sample_MS_v1.mkv 1
AVC040 0 FFV1-HEADER-bits_per_raw_sample_MS_v3.mkv 1
AVC041 0 FFV1-HEADER-num_h_slices_MS_v3.mkv 1
AVC042 0 FFV1-HEADER-num_v_slices_MS_v3.mkv 1
AVC043 0 FFV1-HEADER-quant_table_count_MS_v3.mkv 1
AVC044 0 FFV1-HEADER-QuantizationTable-len_MS_v3.mkv 1
AVC045 0 FFV1-HEADER-QuantizationTable-scale_MS_v3.mkv 1
AVC047 0 FFV1-HEADER-ec_MS_v3.mkv 1
AVC048 0 FFV1-HEADER-intra_MS_v3.mkv 1
AVC049 0 FFV1-HEADER-configuration_record_crc_parity_v3.mkv 1
AVC050 0 FFV1-HEADER-END_MS_v3.mkv 1
AVC051 0 FFV1-SLICE-slice_xywh_h_v3.mkv 1
AVC051 0 FFV1-SLICE-slice_xywh_w_v3.mkv 1
AVC051 0 FFV1-SLICE-slice_xywh_x_v3.mkv 1
AVC051 0 FFV1-SLICE-slice_xywh_y_v3.mkv 1
AVC052 0 FFV1-SLICE-quant_table_index_v3.mkv 1
AVC053 0 FFV1-SLICE-picture_structure_v3.mkv 1
AVC054 0 FFV1-SLICE-FFV1-SLICE-sar_den_v3_den0.mkv 1
AVC054 0 FFV1-SLICE-FFV1-SLICE-sar_den_v3_num0.mkv 1
AVC055 0 FFV1-HEADER-slice_size_v3.mkv 1
AVC056 0 FFV1-SLICE-JUNK_v3.mkv 1
AVC057 0 FFV1-SLICE-slice_crc_parity_v3.mkv 1
AVC058 0 FFV1-SLICE-SliceContent_v3_GR.mkv 1
AVC058 0 FFV1-SLICE-SliceContent_v3_RC.mkv 1
AVC059 0 FFV1-FRAME-END_v3.mkv 1
AVC061 0 MKV-V4.xml 1
AVC061 0 MKV-V4+.mkv 1
AVC062 0 MKV-SEGMENTUID-PRESENT.xml 1
AVC062 0 MKV-SEGMENTUID-PRESENT.mkv 1
AVC063 0 MKV-SEEKHEAD-PRESENT.xml 1
AVC063 0 MKV-SEEKHEAD-PRESENT.mkv 1
AVC064 0 MKV-INTERLACEMENT-CLARITY.xml 1
AVC064 0 MKV-INTERLACEMENT-CLARITY.mkv 1
AVC065 0 MKV-SAMPLE-RANGE-CLARITY.xml 1
AVC065 0 MKV-SAMPLE-RANGE-CLARITY.mkv 1
AVC066 0 MKV-COLOUR-PRIMARY-CLARITY.xml 1
AVC066 0 MKV-COLOUR-PRIMARY-CLARITY.mkv 1
AVC067 0 FFV1-SLICE-error_status_v3.mkv 1
AVC068 0 FFV1-FRAME-key_frame-ISNOTINTRA_MS_v3.mkv 1
References


