DELIVERABLE

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

D8.1R2 Competitive Evaluation Strategy

Revision: Final 1.00, 26 October 2016

Authors:

Nicola Ferro (UNIPD), Erik Buelinckx (KIKIRPA), Boris Doubrov (VeraPDF), Klas Jadeglans (Riksarkivet), Bert Lemmens (PACKED), Jerôme Martinez (MediaConch), Víctor Muṇoz (EasyInnova) Claudio Prandoni (Promoter), Dave Rice (MediaConch), Stefan Rohde-Enslin (SPK), Xavi Tarres (EasyInnova), Erwin Verbruggen (S&V), Benjamin Yousefi (Riksarkivet), Carl Wilson (VeraPDF)

Reviewers:

Peter Pharow

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
P	Public	X
C	Confidential, only for members of the consortium and the Commission Services	

Revision History

Revision	Date	Author	Organisation	Description
0.01	2016-04-13	Nicola Ferro	UNIPD	First skeleton
0.03	2016-04-19	Nicola Ferro	UNIPD	Initial structure
0.10	2016-04-25	Nicola Ferro	UNIPD	Initial draft circulated internally for discussion
0.15	2016-07-15	Nicola Ferro	UNIPD	Updates after consortium feedback and activities kick-off meeting
0.20	2016-10-18	Nicola Ferro	UNIPD	Added content from the working groups on text, image, and audio/video classes
0.30	2016-10-21	Nicola Ferro	UNIPD	Updated version after final working groups meetings
1.00	2016-10-26	Nicola Ferro	UNIPD	Final version after partners feedback

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This deliverable contains original unpublished work except where clearly indicated otherwise.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 Design of the PREFORMA Evaluation Matrix – Release 2 **9**
2.1 Impact on the Challenge . . . 11
2.1.1 Description . . . 11
2.1.2 Items . . . 11
2.2 Technical Approach . . . 16
2.2.1 Description . . . 16
2.2.2 Items . . . 17
2.3 Quality of the Tender . . . 20
2.3.1 Description . . . 20
2.3.2 Items . . . 20
2.4 Costs . . . 21
2.4.1 Description . . . 21
2.4.2 Items . . . 21

3 Tender Procedures **22**
3.1 Workplan . . . 23

4 Testing Framework **24**
4.1 Related Work . . . 24
4.2 Conformance Checking as a Classification Task . . . 24
4.3 Evaluating Conformance Checkers for Digital Preservation . . . 25
4.3.1 Document Collections . . . 26
4.3.2 Ground-Truth . . . 28
4.3.3 Measures . . . 29

5 PREFORMA Testing Classes **31**
5.1 Text Media Type . . . 33
5.2 Image Media Type . . . 49
5.3 Audio-video Media Type . . . 55
5.4 Preparation of the Classes . . . 65
5.4.1 Domain Expert Groups . . . 66
5.5 Preparation of the Ground-Truth . . . 67

6 Testing Workflow **69**
6.1 File Formats . . . 69
6.2 Submission of Runs . . . 70
6.3 Computation of Evaluation Results . . . 71

References **71**

Executive Summary

This second release of D8.1 “Competitive Evaluation Strategy” [Agosti et al., 2014] has a twofold goal:

- In December 2016 there will be the final tender of the PREFORMA project which is aimed at selecting the suppliers which will participate in the “Testing Phase” scheduled from January to June 2017. This deliverable defines the criteria according to which suppliers participating in this tender will be evaluated and compared in order to determine which of them will actually proceed to the “Testing Phase”.
- The “Testing Phase” will evaluate the tool produced by the suppliers on real experimental collections in order to assess their overall quality for conformance checking. This deliverable defines the methodologies and protocols which will be used in this phase to assess the suppliers’ tools.

The document is organized as follows: Section 2 describes the new instantiation of the “PREFORMA Evaluation Matrix” tailored for evaluating the access to the “Testing Phase”; Section 3 introduces the procedures according to which the tender in December 2016 will be managed; Section 4 introduces the framework which will be adopted to evaluate suppliers’ tools during the “Testing Phase”; Section 5 details, for each media type targeted by PREFORMA, the testing classes which will be used; Section 6 describes the practical workflow which will be followed to operate the “Testing Phase”.

1 Introduction

This second release of D8.1 "Competitive Evaluation Strategy" [Agosti et al., 2014] has a twofold goal:

- In December 2016 there will be the final tender of the PREFORMA project which is aimed at selecting the suppliers which will participate in the "Testing Phase" scheduled from January to June 2017. This deliverable defines the criteria according to which suppliers participating in this tender will be evaluated and compared in order to determine which of them will actually proceed to the "Testing Phase".
- The "Testing Phase" will evaluate the tool produced by the suppliers on real experimental collections in order to assess their overall quality for conformance checking. This deliverable defines the methodologies and protocols which will be used in this phase to assess the suppliers' tools.

When it comes to the first goal, this deliverable adopts the same evaluation model as defined in Section 2 of D8.1 "Competitive Evaluation Strategy" [Agosti et al., 2014] but it provides a new instantiation of the "PREFORMA Evaluation Matrix" (Section 3 of the first release of D8.1), which is more suitable for evaluating whether a supplier has to continue or not with the "Testing Phase".

When it comes to the second goal, this deliverable defines a specific evaluation framework [Ferro, 2016], based on the well-known Cranfield paradigm [Cleverdon, 1997] and shared among leading evaluation initiatives worl-wide, such as *Text REtrieval Conference (TREC)*[1] [Harman and Voorhees, 2005] in the United States, *Conference and Labs of the Evaluation Forum (CLEF)*[2] [Ferro, 2014] in Europe, *NII Testbeds and Community for Information access Research (NTCIR)*[3] in Japan and Asia, and *Forum for Information Retrieval Evaluation (FIRE)*[4] in India.

The document is organized as follows: Section 2 describes the new instantiation of the "PREFORMA Evaluation Matrix" tailored for evaluating the access to the "Testing Phase"; Section 3 introduces the procedures according to which the tender in December 2016 will be managed; Section 4 introduces the framework which will be adopted to evaluate suppliers' tools during the "Testing Phase"; Section 5 details, for each media type targeted by PREFORMA, the testing classes which will be used; Section 6 describes the practical workflow which will be followed to operate the "Testing Phase".

2 Design of the PREFORMA Evaluation Matrix – Release 2

As explained in Section 2 of the first release of D8.1 [Agosti et al., 2014], designing the evaluation matrix involves the following steps:

1. defining reviewer types;

[1] http://trec.nist.gov/
[2] http://www.clef-initiative.eu/
[3] http://research.nii.ac.jp/ntcir/index-en.html
[4] http://fire.irsi.res.in/

2. defining categories, their respective weights and the weights of reviewer types within each category;

3. defining items and their respective weights.

We need to consider three reviewer types, which correspond to the three main stakeholders involved in the PREFORMA project, as shown in Figure **??**. They are:

- **Technical Expert**: the proposed solution is analyzed by a technical expert, who is evaluating the solution from the technical point-of-view;
- **Domain Expert**: the proposed solution is analyzed by a domain expert, who verifies if the solution well fits the requirement of the domain where it will be used;
- **External Expert**: the proposed solution is analyzed by an expert external to the PREFORMA consortium, to compensate for any possible biases.

In particular, each supplier solution will be reviewed by the following number of reviewers for each reviewer type:

- **Technical Expert**: 2 reviewers for each proposal, i.e. $rt' = 2$;
- **Domain Expert**: 3 reviewers for each proposal, i.e. $rt'' = 3$;
- **External Expert**: 1 reviewer for each proposal, i.e. $rt''' = 1$;

The following is a detailed description of all the categories and all the items for each category is presented. Each category is specified by:

- name;
- brief description;
- weight of the category;
- weights of the different reviewer types within the category;
- complete list of all the items of the category.
  Each item is specified by:
  - name;
  - brief description;
  - meaning of the scores for that item;
  - weight of the item.

This deliverable keeps the same categories introduced in the first release of D8.1, since they are at the core of the PREFORMA challenge brief [Lemmens, 2014] and tender [Lemmens et al., 2014].

2.1 Impact on the Challenge

2.1.1 Description

The category

> Impact On the Challenge concerns the extent of how well the proposed idea, solution or technology meets the challenge as detailed in the Challenge Brief, and whether it will have the desired impact.
>
> This category pertains all three reviewer types.
>
> The following parameters are valid for the category:
>
> - Category C_1 – IMPACT ON THE CHALLENGE
> - *Category weight*: $\alpha_1 = 35\%$
> This category evaluates the extent to which the proposed solution meets the challenge.
> - *Reviewer Type weights*:
> - *"Technical Expert" reviewer type weight*: $\gamma_1' = 30\%$
> - *"Domain Expert" reviewer type weight*: $\gamma_1'' = 40\%$
> - *"External Expert" reviewer type weight*: $\gamma_1''' = 30\%$

2.1.2 Items

- Item $I_{1,1}$ – **Basic research questions**
  - *description*: This item concerns: (i) how to interpret and implement standard specifications; (ii) how to determine whether a file is what it claims to be; and (iii) how to make OS project sustainable.
    In order to evaluate the aspect (i) the reviewer has to consider if the project establishes a methodology or an objective frame of reference to interpret and implement the standard specifications against the background of the current variations of interpretations and implementations by software vendors and if there is a need to consolidate the diverse implementations or if it is a better to centralize the interpretation to a specific implementation (i.e. promote one interpretation and implementation as the standard).
    In order to evaluate the aspect (ii) the reviewer has to determine whether a file is what it claims to be, i.e., in this context, what makes a file a valid file, or is it conform to the "standard"?
    In order to evaluate the aspect (iii) the reviewer has to consider how the open source project can be developed and sustained in the short and long run and if an open source community can operate as the normative source for the answer to the first and second question.

  - *item weight*: $\beta_{1,1} = 10\%$

- Item $I_{1,2}$ – **Conformance Checker**
  - *description*: This item concerns the conformance checker and it contains some general aspects and some aspects which are specific of the file type the checker is operating with.
    For evaluating the general aspects the reviewer has to consider: (i) if the project develops an open-source conformance checker; (ii) if the conformance checker enables implementation of the OAIS Quality assurance function at Ingest, validating (QA results) the successful transfer of the SIP to the temporary storage area; (iii) if the conformance checker enables implementation of the OAIS Generate AIP function at ingest, transforming one or more SIPs into one or more AIPs that conform to the archive's data formatting standards and documentation standards; (vi) if the conformance checker enables implementation of the OAIS Archival Information Update function at ingest, providing a mechanism for updating (repackaging, transformation) the contents of the archive.
    For evaluating the file type specific aspects the reviewer has to consider: (PDF/A) if basic research activities involve checking for the existence of PDF/A functionalities and if they are implemented in accordance with the specifications for PDF/A; (TIFF) if basic research activities involve checking for the existence of TIFF functionalities and if they are implemented in accordance with the specifications for TIFF; (AV) if basic research activities involve the definition of a profile for an audiovisual preservation file.
    The conformance checker would be used to evaluate and if possible fix a SIP and convert it into an AIP. Practically this would mean: Your repository receives an exotic TIFF or PDF file and convert it into a TIFF/A or PDF/A file, not by transcoding the file, but by stripping/adding/editing information in the header and the structure of the file. This was how we conceived the Generate AIP function in the OAIS model.
    As for evaluating the use of OAIS as a reference framework, "enable" refers meeting the technical requirements for integrating the conformance checker in existing workflows and designing a shell that allows for performing essential tasks that fit the OAIS framework. So for example Generate OAIS requires a machine readable report that can be used by a transcoder to convert the file. Or when small errors are concerned, the fixer module should be able to perform the conversion.
  - *item weight*: $\beta_{1,2} = 20\%$

- Item $I_{1,3}$ – **Reference Implementation**
  - *description*: This item considers many aspects of the reference architecture of the proposed solutions. The aspects that need to be evaluated by the reviewer

are:
(i) healthy ecosystem: the project establish a healthy ecosystem around an open source 'reference' implementation for specific file formats;
(ii) demonstration files: technology providers contribute demonstration files with good and bad samples of the corresponding reference implementation;
(iii) documentation of the source code: technology providers contribute comprehensive documentation of the source code, which allows for automated generation of the internal API of the application;
(iv) documentation of the software: technology providers contribute comprehensive documentation of the conformance checker for developers, such as quick start guide, cookbooks and other tutorials;
(v) online technical support: technology providers ensure online availability at the development platform for technical support to other developers deploying the conformance checker;
(vi) marketing at conferences: technology providers market the reference implementation and conformance checker at conference for professional networks of developers and digital preservationists;
(vii) propose changes and additions: technology providers draft proposals for changes and additions to the standard specifications;
(viii) participate in work-groups: technology providers participate in technical workgroups that maintain a standard specification;
(ix) facilitate OAIS Monitor Designated Communities: the network of common interest enables implementation of the OAIS Monitor Designated Communities function for Preservation Planning, interacting with Archive Consumers and Producers to track changes in their service requirements and available product technologies;
(x) facilitate OAIS Develop Preservation Strategies and Standards: the network of common interest enables implementation of the OAIS Develop Preservation Strategies and Standards function for preservation planning, developing and recommending strategies and standards, and for assessing risks, to enable the Archive to make informed trade-offs as it establishes standards, sets policies, and manages its system infrastructure;
(xi) facilitate OAIS Establishing Standards and Policies: the network of common interest enables implementation of the OAIS Establishing Standards and Policies function by the Administration of the Archive system and maintain them.

  - *item weight*: $\beta_{1,3} = 20\%$

- Item $I_{1,4}$ – **Future/wider challenges/future proof**
  - *description*: The reviewer has to evaluate the potential of the proposal to address future/wider challenges in the area in an innovative way (e.g. by developing or employing novel concepts, approaches, methodologies, tools, or

technologies).
  - *item weight*: $\beta_{1,4} = 5\%$

- Item $I_{1,5}$ – **Commercial feasibility**
  - *description*: The reviewer has to evaluate the extent to which the approach demonstrates commercial feasibility, and whether it is a realistic commercialization plan or route to market; in particular it has to be considered the following aspects: (i) integration with text, image, moving image editors; (ii) integration with digital repositories; (iii) integration with transcoding software; (iv) integration of additional conformance checkers; (v) integration of additional reporters; (vi) providing consulting services; (vii) providing customization services; (viii) providing support services.
  - *item weight*: $\beta_{1,5} = 15\%$

- Item $I_{1,6}$ – **Open source work practices**
  - *description*: The reviewer has to evaluate this item considering that the development of software in open source projects in PREFORMA must utilise effective open source work practices.
    In particular it has to be considered the following aspects: (i) nigthly builds; (ii) open development platform; (iii) issue/bug trackers; (iv) developer communication channels (e.g. use of forums, use of mailing lists for different stakeholder groups (users, developers, etc.) and use of IRC, provision of roadmaps, provision of documentation, provision of easy hacks, etc.).
  - *item weight*: $\beta_{1,6} = 5\%$

- Item $I_{1,7}$ – **Open Source release practice / Delivery and installation**
  - *description*: This category concerns aspects regarding the delivery and installation of a system.
    In particular it has to be considered the following aspects: (i) executable source code: for each executable of developed software that is provided in an open source project, the source code must always be provided for that executable.;
    (ii) instructions for making executables: for each executable of developed software that are provided in an open source project, instructions for how to create the executable from the source code must always be provided;
    (iii) open source tools for making executables: for each executable of developed software that are provided in an open source project at the PREFORMA open source portal, open source tools (provided under any license approved by Open Source Initiative) for creation of the executable from the source code must be provided;
    (iv) executables for multiple platforms: there must always be executables for

several different platforms (at least for: MS Windows 7, Mac OSX, common Linux distributions such as Ubuntu, Fedora, Debian, and Suse).

- *item weight*: $\beta_{1,7} = 5\%$

- Item $I_{1,8}$ – **Open Source interaction practice**
  - *description*: Individuals in companies contracted by PREFORMA will adopt a work-practice which promote a diverse long-term sustainable Open Source community (which have active participants and contributors from several different organisations).
    In particular it has to be considered the following aspects: (i) engage in timely fashion: companies contracted by PREFORMA for development and provision of software and associated digital assets in Open Source projects must be responsive with respect to contributions to the project and are expected to engage in activities in a timely fashion;
    (ii) open collaboration: companies contracted by PREFORMA for development and provision of software and associated digital assets in Open Source projects must be responsive with respect to contributions to the project and are expected to promote an open collaboration and become active community members which adhere to established community values and work-practices;
    (iii) promote external contribution: companies contracted by PREFORMA for development and provision of software and associated digital assets in Open Source projects must be responsive with respect to contributions to the project and are expected to promote external contributions to each Open Source project;
    (iv) contribute to other projects: companies contracted by PREFORMA for development and provision of software and associated digital assets in Open Source projects must be responsive with respect to contributions to the project and are expected to be active contributors in other relevant Open Source projects that are related to the Open Source project for which they are contracted;
    (v) interact with standardisation organisations: the open source projects conducting development of software for PREFORMA must actively engage in interacting with relevant organisations that maintain the standard specifications used by the open project. The aim is to provide feedback, resolve technical issues;
    (vi) interact with software providers: the open source projects conducting development of software for PREFORMA must actively engage in interacting with relevant software providers (i.e. those providers which have developed software used for creation of files in the specific file format checked by the PREFORMA software) for provision of feedback, resolving technical issues, and contribute in a dialogue for improvement of their interpretation of the technical specifications of standards implemented in their software;

(vii) respect of the negotiation protocol.
  - *item weight*: $\beta_{1,8} = 5\%$

- Item $I_{1,9}$ – **Open Source IPR distribution**
  - *description*: This category has to be evaluated by considering the following aspects:
    (i) software and source code: “GPLv3 or later” and “MPLv2 or later”: all software developed during the PREFORMA project must be provided under the two specific open source licenses: “GPLv3 or later” and “MPLv2 or later”;
    (ii) open formats EIFv1.0/open standards: All digital assets developed during the PREFORMA project must be provided in open file formats, i.e. an open standard as defined in the European Interoperability Framework for Pan-European eGovernment Service (version 1.0 2004). This item concerns the degree of proprietary solution, i.e. if the system uses an open standard solution or a proprietary solutions;
    (iii) CC-BYv4.0: all digital assets developed during the PREFORMA project must be provided under the open access license: Creative Commons CC-BY v4.0; (vii) respect of the negotiation protocol.
  - *item weight*: $\beta_{1,9} = 5\%$

- Item $I_{1,10}$ – **Negotiation Protocol**
  - *description*: The extent to which the recommendations expressed in the negotiation protocol have been implemented in the project.
  - *item weight*: $\beta_{1,10} = 10\%$

2.2 Technical Approach

2.2.1 Description

This category regards all the technical aspects concerning the proposal.

The following parameters are valid for the category:

- Category C_2 – TECHNICAL APPROACH
- *Category weight*: $\alpha_2 = 35\%$
  This category evaluates the technical quality of the proposal.
- *Reviewer Type weights*:
  - *“Technical Expert" reviewer type weight*: $\gamma_2' = 50\%$
  - *“Domain Expert" reviewer type weight*: $\gamma_2'' = 20\%$
  - *“External Expert" reviewer type weight*: $\gamma_2''' = 30\%$

2.2.2 Items

- Item $I_{2,1}$ – **Architecture**
  - *description*: This item concerns infrastructural aspects, technical specifications and system features of a system. In particular it has to be evaluated by considering:
    (i) Interoperability: this item concerns the degree at which the solution can interoperate with other components and solutions;
    (ii) Scalability: this item concerns the degree at which the solution is scalable and expansible;
    (iii) Portability: source code must be built for portability between technical deployment platforms (platform independent);
    (iv) Modularity: source code must be built in a modular fashion for improved maintainability;
    (v) Deployment: the Conformance Checker must allow for deployment in the five infrastructures/ environments defined in the Challenge Brief, i.e. PREFORMA website, stand alone, networked, in legacy system and in test environment;
    (vi) Interface: the Conformance Checker must interface with other software systems via APIs;
  - *item weight*: $\beta_{2,1} = 25\%$

- Item $I_{2,2}$ – **Performances and Quality**
  - *description*: The goal of this item is to evaluate the general performances and the quality, which are measured from both an objective and a subjective point of view.
  - *item weight*: $\beta_{2,2} = 10\%$

- Item $I_{2,3}$ – **Shell Services and features**
  - *description*: This item concerns functionalities and services offered by a system. It has to be evaluated by considering:

    (i) checking at creation time: the Shell component of the Conformance Checker must facilitate conformance checking of files at four moment in the life cycle of a digital document, identified in the use cases of the challenge brief, including conformance checking at creation time, transfer time, digitisation time and migration time;
    (ii) checking at transfer time: The Shell component of the Conformance Checker must facilitate conformance checking of files at four moment in the life cycle of a digital document, identified in the use cases of the challenge brief, including conformance checking at creation time, transfer time, digitisation time

and migration time; (iii) checking at digitisation time: The Shell component of the Conformance Checker must facilitate conformance checking of files at four moment in the life cycle of a digital document, identified in the use cases of the challenge brief, including conformance checking at creation time, transfer time, digitisation time and migration time;
(iv) checking at migration time: The Shell component of the Conformance Checker must facilitate conformance checking of files at four moment in the life cycle of a digital document, identified in the use cases of the challenge brief, including conformance checking at creation time, transfer time, digitisation time and migration time;
(v) automated checks: The Shell component of the Conformance Checker must allow for automating the procedures for checking, reporting and fixing preservation file;
(vi) periodical checks: The Shell component of the Conformance Checker must allow for configuring fully automated, periodical checks;
(vii) batch processing: The Shell component of the Conformance Checker must allow for batch processing of extensive file sets;
(viii) addotopma: The Shell component of the Conformance Checker must allow for configuration of additional components in particular implementation checkers, policy checkers and reporters for other preservation file formats that are developed in the PREFORMA ecosystem;
(ix) use by non-expert users: The Shell component of the Conformance Checker must allow for use by non-expert users;
(x) operate without Internet: The Shell component of the Conformance Checker must be operational in a closed zone with no Internet access.

  - *item weight*: $\beta_{2,3} = 25\%$

- Item $I_{2,4}$ – **Implementation Checker Services and features**
  - *description*: This item has to be evaluated following different criteria on the basis of the file type the checker is designed for.
    For text the checker has to test the compliancy of: PDF 1.4 (PDF/A-1) [ISO 19005-1, 2005], PDF 1.7 [ISO 32000-1, 2008], PDF/A-2 [ISO 19005-2, 2011] and PDF/A-3 [ISO 19005-3, 2012].
    For images the checker has to test the compliancy of: TIFF/EP [ISO 12234-2, 2001] and TIFF/IT [ISO 12639, 2004].
    For Audio/video the checker has to test the compliancy of: MKV, OGG, Lossless JPEG2000 [ISO/IEC 15444, 2004], Lossless FFV1 and LPCM [IEC 60958, 2014].
  - *item weight*: $\beta_{2,4} = 10\%$

- Item $I_{2,5}$ – **Policy Checker Services and features**

  - *description*: This item has to be evaluated by considering technical metadata for text, technical metadata for image and technical metadata for av.
  - *item weight*: $\beta_{2,5} = 5\%$

- Item $I_{2,6}$ – **Reporter Services and features**
  - *description*: This item has to be evaluated by considering if the checker produces both machine readable report and human readable report.
    Machine readable report must provide preservation metadata for each file checked and allowing external software agents to further process the file. The machine readable report will be produced using a standard XML format, implemented by all conformance checkers in the PREFORMA ecosystem, which allows the reported module to combine output from multiple checker components in one report.
    Human readable report must provide a human readable report, assessing the preservation status of a batch of files as a whole, reporting to a non-expert audience whether a file is compliant with the standard specifications, and addressing improvements in the creation/digitisation process.
  - *item weight*: $\beta_{2,6} = 10\%$

- Item $I_{2,7}$ – **Metadata fixer Services and features**
  - *description*: This item has to be evaluated by considering if the checker:
    (i) aligns embedded metadata: The Metadata fixer component of the Conformance Checker must allow for performing fully automated fixes of incongruities in the metadata embedded in the file, based on the report of the implementation checker. Such automated fixes may include making embedded technical metadata conform with the properties of video and audio essence contained by the preservation file;
    (ii) essences normalising metadata: The Metadata fixer component of the Conformance Checker must allow for performing fully automated fixes of incongruities in the metadata embedded in the file, based on the report of the implementation checker. Such automated fixes may include normalising embedded administrative metadata about the preservation file.
  - *item weight*: $\beta_{2,7} = 5\%$

- Item $I_{2,8}$ – **Proposed Approach for Phase 3**
  - *description*: how the suppliers propose to address the requirements and the desiderata that have been identified for this phase.
  - *item weight*: $\beta_{2,8} = 10\%$

2.3 Quality of the Tender

2.3.1 Description

This category deal with all the aspects related to project management and how the negotiation protocol has been taken into account.

The following parameters are valid for the category:

- Category C_3 – QUALITY OF THE TENDER
- *Category weight*: $\alpha_3 = 15\%$
  This category deal with all the aspects related to project management and how the negotiation protocol has been taken into account.
- *Reviewer Type weights*:
  - *"Technical Expert" reviewer type weight*: $\gamma_3' = 35\%$
  - *"Domain Expert" reviewer type weight*: $\gamma_3'' = 35\%$
  - *"External Expert" reviewer type weight*: $\gamma_3''' = 30\%$

2.3.2 Items

- Item $I_{3,1}$ – **Project plan**
  - *description*: This item evaluates the extent to which the tender shows a clear plan for the development of a working solution, and whether it is a reasonable plan to finish phase 3 in time. It must be verified if the proposal respects of the negotiation protocol.
  - *item weight*: $\beta_{3,1} = 25\%$

- Item $I_{3,2}$ – **Effectiveness management**
  - *description*: This item evaluates the effectiveness of the management.
  - *item weight*: $\beta_{3,2} = 25\%$

- Item $I_{3,3}$ – **Resource allocation**
  - *description*: The extent to which the tenderer and/or subcontractor appear to have dedicated the resources (e.g. human capital, equipment etc.) necessary to perform the scope of the tender. It must be verified if the proposal respects of the negotiation protocol.
  - *item weight*: $\beta_{3,3} = 25\%$

- Item $I_{3,4}$ – **Risk Assessment / Risk factors**

- *description*: The extent to which crucial risks (technical, commercial and other) to project success appear to be identified, and how effectively these will be managed. this item concerns the riskiness of a system and the acceptance of these risks.
- *item weight*: $\beta_{3,4} = 25\%$

2.4 Costs

2.4.1 Description

This category concerns the financial aspects of a supplier.

- Category C_4 – COSTS
- *Category weight*: $\alpha_4 = 15\%$
  This category concerns the financial aspects of a supplier.
- *Reviewer Type weights*:
  - *"Technical Expert" reviewer type weight*: $\gamma_3' = 35\%$
  - *"Domain Expert" reviewer type weight*: $\gamma_3'' = 35\%$
  - *"External Expert" reviewer type weight*: $\gamma_3''' = 30\%$

2.4.2 Items

- Item $I_{4,1}$ – **Price/cost**
  - *item weight*: $\beta_{4,1} = 100\%$

3 Tender Procedures

This PCP will result in a framework agreement with three phases: *Design, Prototyping, and Scientific Testing*. The framework agreement sets out the conditions (rights and obligations between contracting authority and contractors) for the entire duration of the PCP.

Riksarkivet will sign framework agreements with contractors to provide services for the PCP that will start with Phase 1, in the respective areas of open source development. Following the completion of Phase 1, the contracting authority will make a call for bids for R&D services for Phase 2 from contractors that have successfully completed Phase 1 (that is, provided an approved report from Phase 1). Upon completion of Phase 2, a corresponding call for bids for Phase 3 will take place. The assessment criteria and weighting for Phase 1 is set out in this Invitation to Tender. The criteria and weighting for the awards of contracts for subsequent Phases will be based on these, but may be elaborated or developed in further detail within those frames. (Please see Assessment of Tenders below.)

- **Phase 1, Design**, is intended to demonstrate the feasibility of the proposed concepts for new solutions. The contracts placed for Phase 1 will be for the duration of four months, between November 2014 and February 2015. A budget of maximum 390 000 Euros is available for all selected projects. The number of awarded contracts depends on the price of the tenders, and the required minimum score of the tenders.
- **Phase 2, Prototyping**, is intended to develop prototypes from the more promising concepts delivered by the selected suppliers in Phase 1. Participation in Phase 2 depends upon successful completion of Phase 1, and contracts for this phase will be awarded to Phase 1 contractors selected by the Evaluation Committee. This phase will take place for 22 months, between March 2015 and December 2016, and is subdivided into three distinct stages: a) First prototypes, a phase to take place between March and October 2015, b) Re-design, which is planned to take place between November 2015 and February 2016; and c) Second prototype, which is planned for the period between March and December of 2016. The number of awarded contracts depends on the price of the tenders, and the required minimum score of the tenders.
- In **Phase 3, Evaluation**, the applications will be tested by the memory institutions of the consortium. Contracts for Phase 3 will be awarded to contractors that have successfully completed Phase 2. The testing phase will take place for six months, between January and June of 2017. The number of awarded contracts depends on the price of the tenders, and the required minimum score of the tenders.

The indicative amount for all projects in the prototyping and testing phases is 2 415 000 Euros for all selected projects. The framework agreement sets out the framework

conditions for the entire duration of the PCP, covering phase 1, 2 and 3. It remains binding as long as contractors remain in competition. Tenderers shall therefore in their offer not only state their detailed offer for phase 1, but also state their goals, and outline plans (including price conditions) for Phases 2, and 3, as a path to achieve the overall purpose of the project. The payments are firm and fixed in Euros, i.e. not adjusted for foreign exchange and/or index or in any other way. All prices shall be stated in Euros.

3.1 Workplan

14 October 2016 Template for final release report

14 October 2016 Template for End of Phase Report

31 October 2016 Submission of final release report

31 October 2016 Submission of D8.1

15 November 2016 Submission of End of Phase Report

18 November 2016 Admission to the call for tender for Phase 3

21 November 2016 Launch of the call for tender

4 December 2016 Proposals submission

16 December 2016 Final evaluation reports and communication to suppliers

21 December 2016 Negotiation completed (if needed)

23 December 2016 Award decision

31 December 2016 Signature of the contracts for Phase 3

9 January 2017 Kick-off Phase 3

4 Testing Framework

4.1 Related Work

"Digital preservation is about more than keeping the bits [...] It is about maintaining the semantic meaning of the digital object and its content, about maintaining its provenance and authenticity, about retaining its interrelatedness, and about securing information about the context of its creation and use" [Ross, 2012, p. 45]. Since preservation aims at capturing the very essence of digital objects it is often associated with life cycles [Kowalczyk, 2015], preservation actions, and overall preservation frameworks and there is often the need to evaluate them and choose among them [Becker and Rauber, 2011; Innocenti et al., 2009].

When it comes to preservation frameworks and their evaluation, this paper focuses on a specific step of a more general preservation framework, namely the checking for conformance of document with respect to their reference standards at ingestion time. In particular, the focus of the paper is on how to evaluate tools for carrying out this step, i.e. conformance checkers, and how to create a benchmark for this purpose.

The idea of benchmarking tools for preservation is gaining more and more traction recently [Chanod et al., 2010] and we share a similar approach with [Duretec et al., 2015], who identify the main components of a digital preservation benchmark as:

- *motivating comparison* defines the comparison to be done and the benefits that comparison will bring in terms of the future research agenda;
- *task sample* is a list of tests that the subject, to which a benchmark is applied, is expected to solve;
- *performance measures* are qualitative or quantitative measurements taken by a human or a machine to calculate how fit the subject is for the task.

4.2 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 can model the conformance checking process as a classification task [Alpaydin, 2014], 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.

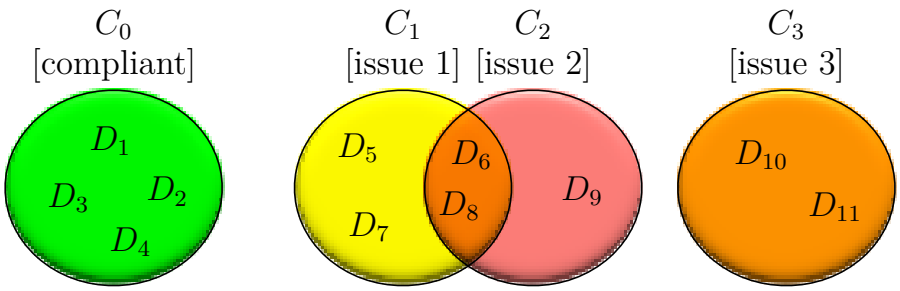

Figure 1: Conformance checking as a classification task.

One of the challenges we have to face is how to determine the list of classes for each the media types targeted by *PREservation FORMATs for culture information/e-archives (PREFORMA)*. Domain experts – both from memory institutions and with technical skills on each specific media type – play a central role in this respect, since they can point out known validation issues, potential validation issues, preservation issues also related to policies of memory institutions, and so on.

One critical aspect in determining such classes is related to their cardinality and granularity. Producing hundreds and hundreds of classes for each media type may be tempting, if you consider this as an indicator of exhaustiveness, but it risks to be harmful in practice, since you may simply ask too much to a conformance checker and you may focus on too tiny or almost irrelevant compliance violations. Therefore, the class creation process must be conducted in an iterative way and domain experts need to work in panels, where they revise and refine each other proposals trying to find the right balance between exhaustiveness and usefulness.

In order to provide an additional degree of flexibility to conformance checking, and its evaluation, we plan to also attach a *severity* to each class since some issues are errors, some others are warnings, some others are mis-conformances to policies and best practices, as it is also shown by the different classes color in Figure 1. If further analysis and requirements will support it, this could even be turned into a full *meta-classification* of the identified classes, in order to allow us to group them on the basis of their semantics and relationships and, for example, to express progressive levels of conformance, like core, intermediate and full.

4.3 Evaluating Conformance Checkers for Digital Preservation

In order to evaluate conformance checkers, we will rely on the Cranfield paradigm [Cleverdon, 1997], which makes use of experimental collections $\mathcal{C} = (D, T, GT)$, 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, as discussed in Section 4.2, the gathered documents, as described in Section 4.3.1, and the ground-truth, as presented in Section 4.3.2; the *performance measures* are described in Section 4.3.3.

4.3.1 Document Collections

The preparation of the collection of documents to be used for assessing the performances of a conformance checker is a critical task that needs to be driven by domain experts. We need to gather a huge sample (ten thousands) for each media type (text, image, audio) from the memory institutions participating in PREFORMA, from the suppliers which are developing the conformance checker tools, and from the open source community at large, which is being built around the PREFORMA effort.

Documents must be representative of the different classes C_i we need to evaluate conformance checkers against. In particular, we cannot have empty classes, i.e. classes for which there is no document in the experimental collection, and the cardinality of each class, i.e. the number of documents in the collection belonging to that class, should make sense from two points of view. Firstly, it should have a size, relative to the other classes, which is proportional to the frequency of the issue represented by the class in real world settings; in other terms, there are issues that happen more frequently and there are issues which are more rare and this should be reflected in the cardinality of the corresponding classes, in order to confront conformance checkers with realistic settings. Secondly, we should pay attention to not introduce any bias in the evaluation measurement and process due to an uncontrolled and excessive discrepancy in the cardinality of the classes.

Figure 2 shows the main data set which will be used and made available during the lifetime of the project [Elfner and Justrell, 2014]. The main distinction is between:

- *training dataset*: aimed at driving and facilitating the design and development of supplier systems, i.e. conformance checkers, as well as show casing their functionalities.
- *test dataset*: aimed at evaluating and testing the supplier systems in order to score and subsequently select the best of them.

Test and training datasets are kept as two distinct datasets, i.e. there is no intersection, in order to avoid overfitting supplier system on datasets and to ensure fair and unbiased assessment of them.

Both training and test dataset will be associated with ground-truth specifying the correct labels for the documents in the dataset but the ground-truth associated with the test data

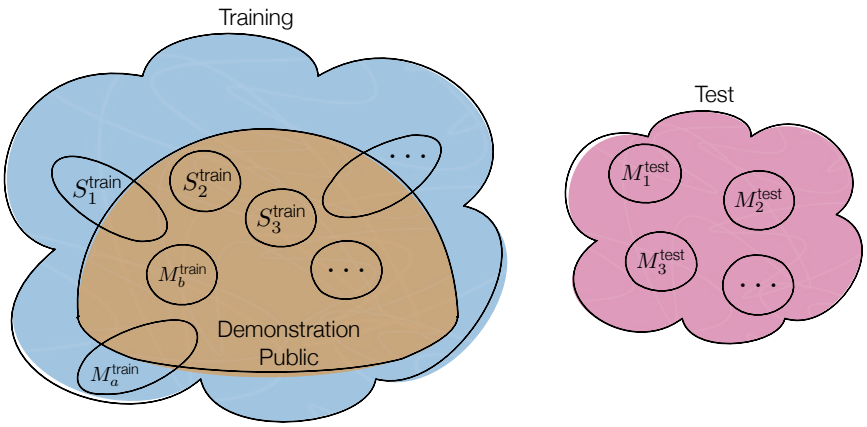

Figure 2: PREFORMA document collections.

set will not be shared ahead, because it is needed for carrying out the final testing phase in an unbiased way.

More in detail, the test dataset is constituted by representative test data M_j^{test} provided by memory institutions that can be either partners of the PREFORMA consortium or members of the PREFORMA network of memory institutions. During the execution of the PREFORMA project, this dataset is private and it will be shared only within the consortium to test the supplier systems. After the end of the PREFORMA project, memory institutions may decide to make (part of) it public to favour the PREFORMA ecosystem and open source community.

The training dataset is constituted by: (i) representative training data M_k^{train} provided by memory institutions that can be either partners of the PREFORMA consortium or members of the PREFORMA network of memory institutions; (ii) representative training data S_k^{train} provided by the suppliers participating in the project.

The training dataset is constituted by two parts: a *demonstration* one, which is public and serves the purpose of show casing the suppliers systems both to the other suppliers and to the memory institutions; a *private* part, which is used internally by each supplier for designing, developing, and testing its own system.

Data provided by memory institutions and suppliers which are in the demonstration dataset are accessible and shared also with the other suppliers participating in the project, besides the general public. The purpose of the demonstration dataset is to trigger and facilitate the growth and development of the PREFORMA ecosystem, the open source community, the communication with standardization bodies and, if properly fed, will represent also a strategic asset for suppliers in order to sustain their own business plans.

An orthogonal distinction on the datasets is between *synthetic* and *real* data. The former are data created with the specific purpose of pinpointing some specific compliance

Class C_i		Ground-Truth	
		Positive	Negative
Conformance Checker	Positive	True Positive (TP_i)	False Positive (FP_i)
	Negative	False Negative (FN_i)	True Negative (TN_i)

Figure 3: Confusion matrix for the evaluation of conformance checkers for each class C_i.

problem or critical issue for a given preservation format, as proposed also by [Becker and Duretec, 2013]. The latter are data actually managed by memory institutions for their preservation duties. It is intended that both the training and the test datasets will be comprised by both synthetic and real data.

4.3.2 Ground-Truth

As it is well known [Sanderson, 2010], ground-truth creation is an extremely demanding activity since it requires a great amount of human effort to be conducted. For this reason, a lot of research concentrated on how to reduce the burden of ground-truth creation ranging from the utopian attempt to eliminate assessments at all [Soboroff et al., 2001] to crowdsourcing [Alonso, 2013; Lease and Yilmaz, 2013].

Unfortunately, in the context of PREFORMA, crowdsourcing it is not a viable option since real domain experts are needed to carefully judge the compliance of a document to its reference standard.

Two interesting questions will arise during ground-truth creation in PREFORMA. The first issue is that, to assess the compliance of a document, domain experts will probably also use some of the already existing tools and this may introduce circularity and bias. The second issue is to understand the problem of inter-assessor agreement and see whether on this highly specialised task it will have similar ratios as those for ad-hoc retrieval [Voorhees, 2000], i.e. in the range 30%–50%, or whether discrepancies from previously known tasks will arise.

The above issues apply in the case of the *real* data while *synthetic* data help mitigating the burden of ground-truth creation, because each synthetic document is purposefully created for testing one or more issues in complying to a standard and it is therefore automatically labeled since its creation.

4.3.3 Measures

Evaluating conformance checkers is not a binary process, i.e. it is not like going through a long check-list and if any of the items in the list is missing or incorrect, the conformance checker is rejected. The evaluation we foresee is more flexible and we aim at quantifying the extent a conformance checker is able to spot deviations from its reference standard.

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 3.

Recall from Section 4.2 and Figure 1 that each class C_i represents a possible misconformance 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 Positve (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.

Note that what we mean by the confusion matrix of Figure 3 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.

This duality between the harshness of FNs and FPs resembles a similar duality between spam and ham misclassification [Cormack and Lynam, 2005], where spam misclassification annoys the user and may cause the user to overlook important messages while ham misclassification inconveniences the user and risks loss of important messages.

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

$$\text{Accuracy}_i = \frac{|TP_i| + |TN_i|}{|TP_i| + |TN_i| + |FP_i| + |FN_i|} \tag{1}$$

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

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

- *logistic average misclassification rate (LAM)*: is the geometric mean of the *odds* of compliance and not-compliance misclassification, converted back to a proportion [Cormack and Lynam, 2005; Smucker et al., 2013]. This measure imposes no a priori relative importance on compliance and not-compliance misclassification, and rewards equally a fixed-factor improvement in the odds of either.

$$\text{LAM}_i = \text{logit}^{-1}\left(\frac{\text{logit}\,(fpr) + \text{logit}\,(fnr)}{2}\right) \tag{3}$$

  where $fpr = \frac{|FP_i|}{|FP_i|+|TN_i|}$ is the *false-positive rate*, $fnr = \frac{|FN_i|}{|FN_i|+|TP_i|}$ is the *false-negative rate*, and the logit transformations are given by $\text{logit}(x) = \ln\frac{x}{1-x}$ and $\text{logit}^{-1}(x) = \frac{\mathrm{e}^x}{1+\mathrm{e}^x}$.

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 4.2, 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:

$$\begin{aligned} \text{Consistency} &= 1 - \frac{\sum_{i=1}^{N} |(TP_0 \cup FP_0) \cap (TP_i \cup FP_i)|}{\sum_{i=1}^{N} |(TP_i \cup FP_i)|} \\ &= 1 - \frac{\sum_{i=1}^{N} |C_0 \cap C_i|}{\sum_{i=1}^{N} |C_i|} \end{aligned} \tag{4}$$

where N is the total number of classes, excluded C_0. 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.

Figure 4 shows some relevant cases for consistency: when there is no intersection between C_0 and the other classes then Consistency $= 1$ (Figure 4.a); on the other hand, in the extreme case of complete overall between C_0 and the other classes, i.e. when all the documents are assigned to all the classes, Consistency $= 0$ (Figure 4.c); in the other cases, when some overlap exists, consistency is in the range $(0, 1)$ (Figure 4.b).

5 PREFORMA Testing Classes

The media types addressed by PREFORMA are: (i) *electronic documents* for establishing a reference implementation for PDF/A [ISO 19005-1, 2005; ISO 19005-2, 2011; ISO 19005-3, 2012]; (ii) *images* for establishing a reference implementation for uncompressed TIFF [ISO 12234-2, 2001; ISO 12639, 2004]; and, (iii) *audio-video* for establishing a reference implementation for an audiovisual preservation file, using FFV1[5] for encoding video or moving images, uncompressed LPCM [IEC 60958, 2014] for encoding sound and MKV[6] for wrapping audio- and video-streams in one file.

[5] http://www.ffmpeg.org/~michael/ffv1.html
[6] http://www.matroska.org/

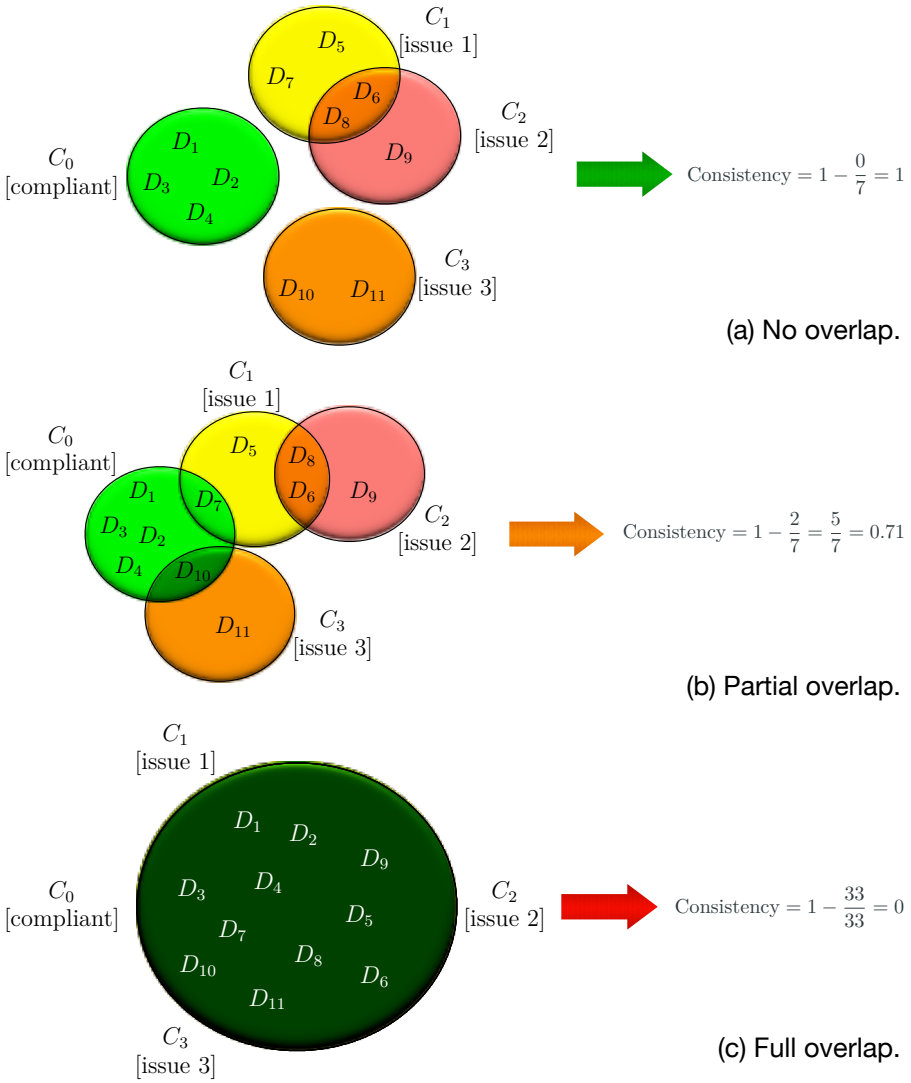

Figure 4: Different cases for consistency: (a) no overlap between C_0 and the other classes; (b) partial overlap between C_0 and the other classes; (c) complete overlap between C_0 and the other classes.

For each class it is specified:

- a unique identifier;
- a short name (in bold);
- a brief description;
- the type of the class – either Conformance or Policy – to indicate whether it is aimed at evaluating a conformance checker or a policy cheker;
- the severity of the class in the range $[1, 5]$, i.e. the extent to which the conformance/policy issue the class is about is crucial. 1 indicates the minimum severity whereas 5 indicates the maximum severity.

For each of these media types, the following sections detail the specific classes against which suppliers' tools will be evaluated. The last two sections describe how the classes have been collected and how the ground-truth has been created.

5.1 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 $TC001$ – **Annotation FileAttachment**
  - *description*: Has File attached to the PDF/A document.
  - *PDF/A version*: 1
  - *conformance level*: ba
  - *class type*: Conformance
  - *severity*: 5
- 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 Annnotation 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 $TC011$ – **Image Alternative/Proxy**
  - *description*: Has alternative/proxy versions of an image. Usually a preview version, in lower quality, of an image.
  - *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 $TC015$ – **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
- 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 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**
  - *description*: Has a go-to-3D-view action. Identifies a 3D annotation and specifies a view for the annotation to use (PDF reference, sec. 12.6.4.15). Introduced in PDF 1.6.
  - *PDF/A version*: 1-3
  - *conformance level*: bua
  - *class type*: Conformance
  - *severity*: 5
- Class $TC034$ – **Named Action**
  - *description*: Has ANY other Named Action besides NextPage, PrevPage, FirstPage, LastPage.
  - *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*: Conformace
  - *severity*: 5
- Class $TC039$ – **File header**
  - *description*: Has
  - *PDF/A version*: 2-3
  - *conformance level*: bua
  - *class type*: Conformace
  - *severity*: 5
- Class $TC040$ – **Keyword spacings**
  - *description*: Has spacings around keywords 'obj', 'endobj', 'stream', 'endstream', 'xref'.
  - *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 definied 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 $TC057$ – **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
- 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**
  - *description*: Does not have font metrics (see https://partners.adobe.com/public/developer/en/font/500
  - *PDF/A version*: 1-3
  - *conformance level*: bua
  - *class type*: Conformance
  - *severity*: 5
- 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
- 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 $TC063$ – **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
- 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 $TC072$ – **Attached documet is XML.**
  - *description*: Does not have XML file attached.
  - *PDF/A version*: 3
  - *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 $TC076$ – **Presence of Javascript**
  - *description*: Does not have have javascript
  - *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
- Class $TC087$ – **Digital signature should be present**
  - *description*: The document is not digitially signed
  - *PDF/A version*: Any
  - *conformance level*: Any
  - *class type*: Policy
  - *severity*: 5
- 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

5.2 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 BitsPerSample 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 (NewSubFileType 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 StardardID**
  - *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, JPEGInterchangeFormatLngth, JPEGRestartInterval, JPEGLosslessPredictors, JPEGPointTransforms, JPEGQTables, JPEGDCTables, JPEGACTables, YCbCrCoefficients, YCbCrSubSampling, YCbCrPositioning, ReferenceBlackWhite, CFARepeatPatternDim, CFAPattern, Interlace, CompressedBitsPerPixel, FocalPlaneXResolution, FocalPlaneYResolution, ImageSourceData)

  - *class type*: Policy
  - *severity*: 5
- Class $IC042$ – **Mandatory TIA tags**
  - *description*: Tiff without mandatory TI-A 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
- 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

5.3 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 $AVC005$ – **Elements follow maxOccurs**
  - *description*: Verify maxOccurs of EBML Elements [EBML/EBML-ELEMENT-NONMULTIPLES]
  - *class type*: Conformance
  - *severity*: 1
- Class $AVC006$ – **Elements follow minOccurs**
  - *description*: Verify minOccurs of EBML Elements [EBML/EBML-ELEMENT-CONTAINS-MANDATES]
  - *class type*: Conformance
  - *severity*: 1
- Class $AVC007$ – **EBMLMaxIDLength valid**
  - *description*: EBMLMaxIDLength must be in valid range. [EBML/EBML-VALID-MAXID]
  - *class type*: Conformance
  - *severity*: 1
- Class $AVC008$ – **EBMLMaxSizeLength valid**
  - *description*: EBMLMaxSizeLength must be in valid range. [EBML/EBML-VALID-MAXSIZE]
  - *class type*: Conformance
  - *severity*: 1
- 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

- 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
- Class $AVC011$ – **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 $AVC012$ – **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 $AVC013$ – **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
- Class $AVC014$ – **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
- Class $AVC015$ – **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 $AVC016$ – **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
- Class $AVC017$ – **Element Size Unknown**
  - *description*: only Master Elements may be unknown size [EBML/EBML-ELEM-SIZE-UNK]
  - *class type*: Conformance
  - *severity*: 1
- Class $AVC018$ – **Element Data within Size Limits**
  - *description*: test EBML Schema size restrictions per element [EBML/EBML-WITHIN-SIZE-LIMIT]
  - *class type*: Conformance
  - *severity*: 1
- Class $AVC019$ – **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 $AVC020$ – **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 $AVC021$ – **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 $AVC022$ – **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 $AVC023$ – **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
- Class $AVC024$ – **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 $AVC025$ – **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 $AVC026$ – **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
- 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
- Class $AVC028$ – **EBML Element adhers to size restrictions**
  - *description*: The length of the Element falls within the permited range of the optional size declaration. [EBML/EBML-ELEMENT-IN-SIZE-RANGE]
  - *class type*: Conformance
  - *severity*: 1
- Class $AVC029$ – **EBML Element adhers to range restrictions**
  - *description*: The value of the Element falls within the permited range. [EBML/EBML-ELEMENT-VALID-RANGE]
  - *class type*: Conformance
  - *severity*: 1
- Class $AVC030$ – **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 $AVC031$ – **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
- Class $AVC032$ – **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 $AVC033$ – **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 $AVC034$ – **version**
  - *description*: version 0, 1 or 3" Maximum known version is 3 [FFV1/FFV1-HEADER-version]
  - *class type*: Conformance
  - *severity*: 1
- Class $AVC035$ – **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 $AVC036$ – **micro_version 2**
  - *description*: For version 3, micro_version is 4 [FFV1/FFV1-HEADER-micro_version]
  - *class type*: Conformance
  - *severity*: 1
- Class $AVC037$ – **coder_type**
  - *description*: 0 (Golomb Rice), 1 (Range coder) [FFV1/FFV1-HEADER-coder_type]

  - *class type*: Conformance
  - *severity*: 1
- Class $AVC038$ – **state_transition_delta**
  - *description*: (To be defined), FFV1 [FFV1/FFV1-HEADER-state_transition_delta]
  - *class type*: Conformance
  - *severity*: 1
- Class $AVC039$ – **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 $AVC040$ – **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 $AVC046$ – **initial_state_delta**
  - *description*: initial_state_deltas incoherency [FFV1/FFV1-HEADER-initial_state_delta]
  - *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 $AVC049$ – **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 $AVC050$ – **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 $AVC051$ – **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 $AVC052$ – **quant_table_index**
  - *description*: quant_table_index incoherency [FFV1/FFV1-SLICE-quant_table_index]
  - *class type*: Conformance
  - *severity*: 1

- Class $AVC053$ – **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 $AVC054$ – **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 $AVC055$ – **slice_size**
  - *description*: slice_size is bigger than frame size [FFV1/FFV1-SLICE-slice_size]
  - *class type*: Conformance
  - *severity*: 1
- Class $AVC056$ – **error_status**
  - *description*: 0(no error), 1(slice contained a correctable error) [FFV1/FFV1-SLICE-crc_parity]
  - *class type*: Conformance
  - *severity*: 1
- Class $AVC057$ – **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-SLICE-crc_parity]
  - *class type*: Conformance
  - *severity*: 1
- Class $AVC058$ – **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 $AVC059$ – **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 $AVC060$ – **PCM is valid**
  - *description*: some data is there [FFV1/PCM-IS-VALID]

  - *class type*: Conformance
  - *severity*: 1
- Class $AVC061$ – **Matroska version 4 or greater?**
  - *description*: Is MKV at least version 4 [Matroska/MKV-V4+]
  - *class type*: Policy
  - *severity*: 1
- Class $AVC062$ – **SegmentUID is present?**
  - *description*: A SegmentUID Element is stored. [Matroska/SEGMENTUID-PRESENT]
  - *class type*: Policy
  - *severity*: 1
- Class $AVC063$ – **SeekHead is present?**
  - *description*: A SeekHead Element is stored. [Matroska/SEEKHEAD-PRESENT]
  - *class type*: Policy
  - *severity*: 1
- Class $AVC064$ – **Interlaced video is clarified?**
  - *description*: Interlacement is set specifically even if unknown. [Matroska/INTERLACEMENT-CLARITY]
  - *class type*: Policy
  - *severity*: 1
- Class $AVC065$ – **Video Sample Range is clarified?**
  - *description*: Sample range is set specifically even if unknown. [Matroska/SAMPLE-RANGE-CLARITY]
  - *class type*: Policy
  - *severity*: 1
- Class $AVC066$ – **Video Colour Primary is clarified?**
  - *description*: Colour primary is set specifically even if unknown. [Matroska/COLOUR-PRIMARY-CLARITY]
  - *class type*: Policy
  - *severity*: 1
- Class $AVC067$ – **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 $AVC068$ – **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 $AVC069$ – **No junk data within Matroska**
  - *description*: All Master Elements only contain Elements. [Matroska/NO-JUNK-IN-MATROSKA]
  - *class type*: Policy
  - *severity*: 1

5.4 Preparation of the Classes

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;
- 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.

In order to collect and prepare the classes, three separate forms have been setup, as shown in Figure 5:

- text media type:
  - form: http://tinyurl.com/preforma-text-classes
  - sheet: http://tinyurl.com/preforma-text-classes-sheet
- image media type:
  - form: http://tinyurl.com/preforma-image-classes
  - sheet: http://tinyurl.com/preforma-image-classes-sheet
- audio-video media type:
  - form: http://tinyurl.com/preforma-av-classes
  - sheet: http://tinyurl.com/preforma-av-classes-sheet

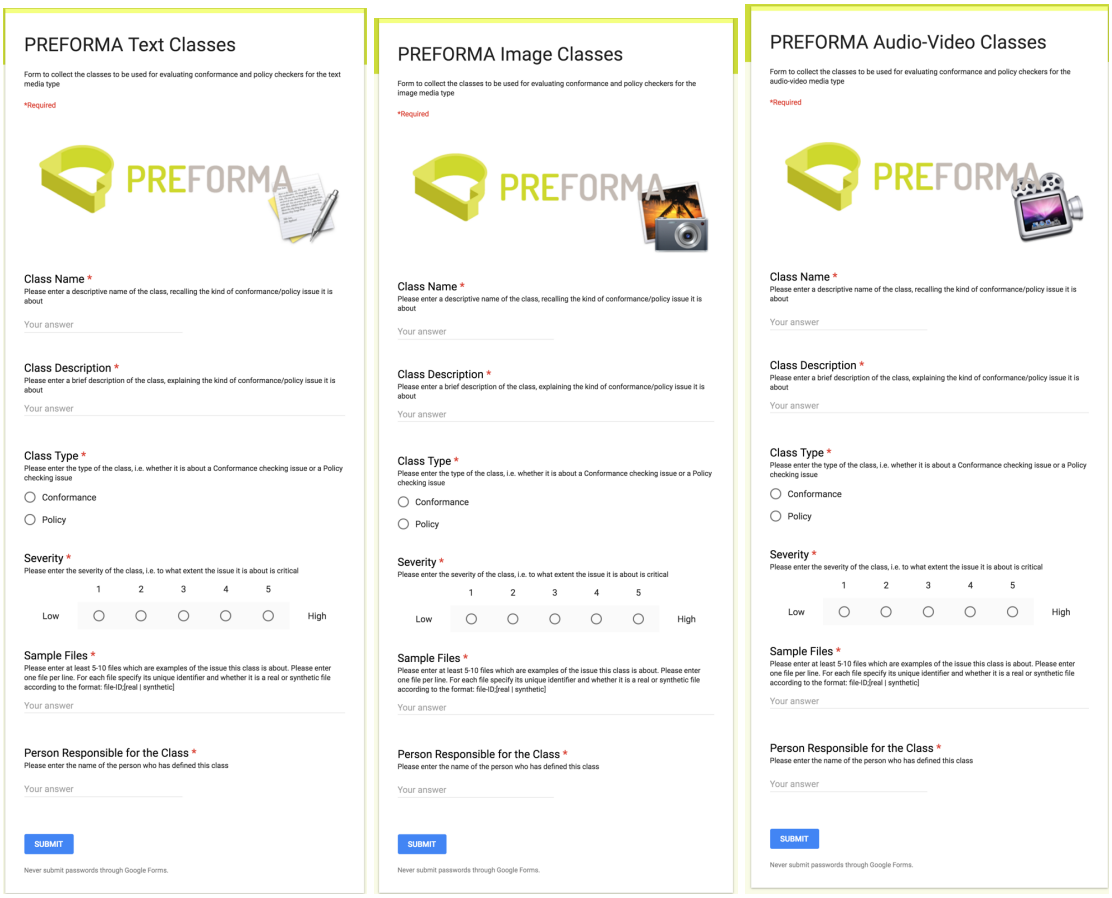

Figure 5: Forms for collecting the classes for the text, image, and audio-video media types.

These forms ask for the same information reported in the previous sections plus an additional list of at least 5-10 files which are true positives for each class (pathnames into the PREFORMA Vault are used as file identifiers). This latter information ensures that each class is based on concrete use cases and that there are actual examples of the issues represented in that class. Moreover, this information represents also a support to ground-truth creation.

The forms were backed by shared Google sheets that allowed each domain expert group to periodically discuss the entered classes and refine them until a final version has been reached.

5.4.1 Domain Expert Groups

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 Muṇoz
  - Bengt Neiss
  - Peter Pharow
  - Stefan Rohde-Enslin
  - Xavi Tarres
  - Christian Weigel
- audio-video media type:
  - Anna Kasimati
  - Uwe Kühhirt
  - Bert Lemmens
  - Emanuel Lorrain
  - Bengt Neiss
  - Peter Pharow
  - Dave Rice
  - Erwin Verbruggen
  - Christian Weigel

5.5 Preparation of the Ground-Truth

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

- 1 evaluation expert;

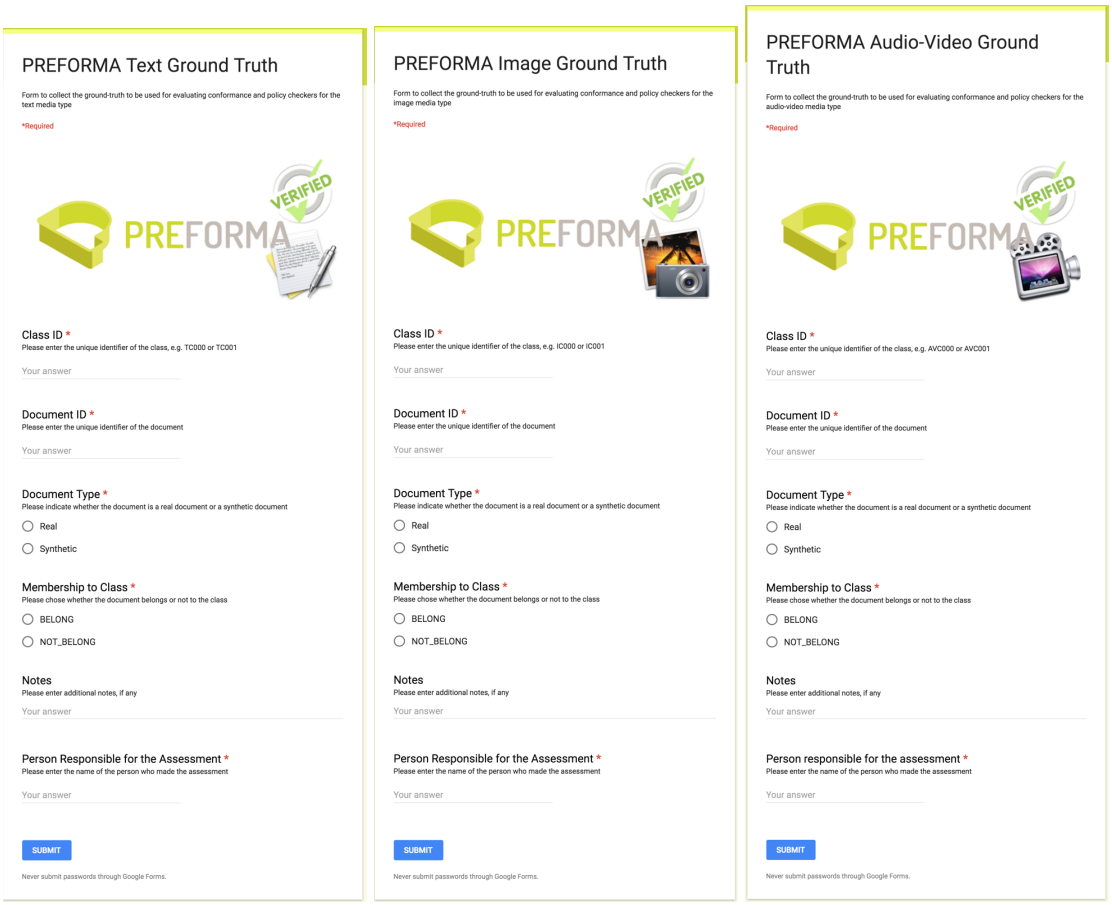

Figure 6: Forms for collecting the ground truth for the text, image, and audio-video media types.

- 3 experts from memory institutions.

Note that, differently from the case of classes creation, the domain expert assessor group does not comprise any member from the suppliers in order to avoid any bias in their evaluation.

In order to collect and prepare the classes, three separate forms have been setup, as shown in Figure 6:

- text media type:
  - form: http://tinyurl.com/preforma-text-ground-truth
  - sheet: http://tinyurl.com/preforma-text-gt-sheet
- image media type:
  - form: http://tinyurl.com/preforma-image-ground-truth
  - sheet: http://tinyurl.com/preforma-image-gt-sheet

- audio-video media type:
  - form: http://tinyurl.com/preforma-av-ground-truth
  - sheet: http://tinyurl.com/preforma-av-gt-sheet

These forms ask for the information needed to determine which documents belong to which classes. In particular, for each pair (class, document), referenced by means of their unique identifiers, the form asks whether the document belongs or not to the class. It collects also a couple of additional information: whether the assessed document is a real or synthetic document and any additional information the assessor may wish to provide.

6 Testing Workflow

The "Testing Phase" will proceed by exchanging textual file between PREFORMA suppliers and the PREFORMA consortium as follows:

- the PREFORMA suppliers will upload their system *runs* to a dedicated repository;
- the PREFORMA consortium will gather those runs and will compute the performance figures for each of them;
- the PREFORMA consortium will upload the performance figures to a dedicated repository;
- PREFORMA suppliers will download their performance figures from a dedicated repository.

6.1 File Formats

We will rely on the standard `trec_eval`[7] textual format adopted in the TREC evaluation campaigns for both ground truth and system runs. In both cases each file is a plain text file where each line of the file is constituted by a set of field separated by a `tab` character

Class ID	0	Document ID	0 not in class 1 in class

Figure 7: Format for each line of a ground-truth file.

Figure 7 shows the formal of a ground-truth file, which indicates for each class which documents belong or not to that class:

- the first field is the unique identifier of a class;

[7] http://trec.nist.gov/trec_eval/

- the second field is ignored and contains the fixed value `Q0`;
- the third field contains the unique identifier of a document;
- the fourth field contains `0` if the document does not belong to the class (true negative) or `1` if it belongs to the class (true positive).

Class ID	Q0	Document ID	0	Score	Run ID

Figure 8: Format for each line of a system run file.

Figure 8 shows the formal of a run file, which indicates for each document in the test set to which class it has been attributed by a supplier system:

- the first field is the unique identifier of a class;
- the second field is ignored and contains the fixed value `Q0`;
- the third field contains the unique identifier of a document;
- the fourth field is ignored and contains the fixed value `0`;
- the fifth field contains the score according to which a system puts a document in a class. This is typically `1` but, if a supplier tool uses some probabilistic techniques, it may be a different value representing the probability for the document of belonging to that class.
- the sixth field contains the unique identifier of the system run.

6.2 Submission of Runs

Each supplier will have a dedicated space in the PREFORMA Vault with two sub-folders, one for uploading their runs and another one for downloading the performance figures.

Each run file must be names according to the following format:

`yyyymmdd_supplierID_runTag.txt`

where

- `yyyymmdd`: it is the date of creation of the run in the ISO 8601 [ISO 8601, 2004] format;
- `supplierID`: it is a unique identifier assigned to each supplier by the PREFORMA consortium;
- `runTag`: a short (and unique if more submission are made in the same day) tag for that run.

Note that the file name without the `.txt` extension has to be used as content of the "Run ID" field in the run format of Figure 8.

Each run file must be complemented with an additional file named

`yyyymmdd_supplierID_runTag.meta.txt`

where a free text description of the run is provided. This information is essential to keep trace of the differences between the various runs submitted by a provider and it will be used for the preparation of Deliverable 8.6 "Testing Report" which will summarize the outcomes of the "Testing Phase".

6.3 Computation of Evaluation Results

During the "Testing Phase", once a week, the PREFORMA consortium will compute the performance figures for the submitted runs and make them available to the suppliers in their dedicated folder in the PREFORMA Vault.

Each performance file will be named as:

`yyyymmdd_supplierID_runTag.results.txt`

References

Agosti, M., Ferro, N., Lemmens, B., and Silvello, G. (2014). Deliverable D8.1 – Competitive Evaluation Strategy. PREFORMA PCP Project, EU 7FP, Contract N. 619568. http://www.digitalmeetsculture.net/wp-content/uploads/2014/12/PREFORMA_D8.1_Competitive-evaluation-strategy_v1.0_no-appendix.pdf.

Alonso, O. (2013). Implementing crowdsourcing-based relevance experimentation: an industrial perspective. *Information Retrieval*, 16(2):101–120.

Alpaydin, E. (2014). *Introduction to Machine Learning*. The MIT Press, Cambridge (MA), USA.

Amigó, E., Gonzalo, J., Artiles, J., and Verdejo, M. F. (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4):461–486.

Amigó, E., Gonzalo, J., and Verdejo, M. F. (2013). A General Evaluation Measure for Document Organization Tasks. In Jones, G. J. F., Sheridan, P., Kelly, D., de Rijke, M., and Sakai, T., editors, *Proc. 36th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2013)*, pages 643–652. ACM Press, New York, USA.

Becker, C. and Duretec, K. (2013). Free Benchmark Corpora for Preservation Experiments: Using Model-Driven Engineering to Generate Data Sets. In Downie, J. S., McDonald, R. H., Cole, T. W., Sanderson, R., and Shipman, F., editors, *Proc. 13th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2013)*, pages 349–358. ACM Press, New York, USA.

Becker, C. and Rauber, A. (2011). Decision Criteria in Digital Preservation: What to Measure and How. *Journal of the American Society for Information Science and Technology (JASIST)*, 62(6):1009–1028.

Chanod, J.-P., Dobreva, M., Rauber, A., Ross, S., and Casarosa, V. (2010). Issues in Digital Preservation: Towards a New Research Agenda. In Chanod, J.-P., Dobreva, M., Rauber, A., and Ross, S., editors, *Report from Dagstuhl Seminar 10291: Automation in Digital Preservation*, Dagstuhl Reports, pages 1–14. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Germany.

Cleverdon, C. W. (1997). The Cranfield Tests on Index Languages Devices. In Spärck Jones, K. and Willett, P., editors, *Readings in Information Retrieval*, pages 47–60. Morgan Kaufmann Publisher, Inc., San Francisco, CA, USA.

Cormack, G. and Lynam, T. (2005). TREC 2005 Spam Track Overview. In Voorhees, E. M. and Buckland, L. P., editors, *The Fourteenth Text REtrieval Conference Proceedings (TREC 2005)*. National Institute of Standards and Technology (NIST), Special Publication 500-266, Washington, USA.

Duretec, K., Kulmukhametov, A., Rauber, A., and Becker, C. (2015). Benchmarks for Digital Preservation Tools. In *Proc. 11th International Conference on Preservation of Digital Objects (iPRES 2015)*.

Elfner, P. and Justrell, B. (2014). Deliverable D2.1 – Overall Roadmap. PREFORMA PCP Project, EU 7FP, Contract N. 619568. http://www.digitalmeetsculture.net/wp-content/uploads/2014/05/PREFORMA_D2.1_Overall-Roadmap_v2.5.pdf.

Fawcett, T. (2006). An Introduction to ROC Analysis. *Pattern Recognition Letters*, 27(8):861–874.

Ferri, C., Hernández-Orallo, J., and Modroiu, R. (2009). An Experimental Comparison of Performance Measures for Classification. *Pattern Recognition Letters*, 30(1):27–38.

Ferro, N. (2014). CLEF 15th Birthday: Past, Present, and Future. *SIGIR Forum*, 48(2):31–55.

Ferro, N. (2016). Proposal for an Evaluation Framework for Compliance Checkers for Long-term Digital Preservation. In Marinai, S., Bertini, M., Orio, N., and Ferilli, S., editors, *Proc. 12th Italian Research Conference on Digital Libraries (IRCDL 2016)*.

Communications in Computer and Information Science (CCIS), Springer, Heidelberg, Germany.

Harman, D. K. and Voorhees, E. M., editors (2005). *TREC. Experiment and Evaluation in Information Retrieval*. MIT Press, Cambridge (MA), USA.

IEC 60958 (2014). Digital audio interface - Part 1: General. Standard IEC 60958-1 Ed. 3.1 b:2014.

Innocenti, P., Ross, S., Maceviciute, E., Wilson, T., Ludwig, J., and Pempe, W. (2009). Assessing Digital Preservation Frameworks: The Approach of the SHAMAN Project. In Spyratos, N., Kapetanios, E., and Traina, A., editors, *Proc. ACM International Conference on Management of Emergent Digital EcoSystems (MEDES 2009)*, pages 412–416. ACM Press, New York, USA.

ISO 12234-2 (2001). Electronic still-picture imaging – Removable memory – Part 2: TIFF/EP image data format. Recommendation ISO 12234-2:2001.

ISO 12639 (2004). Graphic technology – Prepress digital data exchange – Tag image file format for image technology (TIFF/IT). Recommendation ISO 12639:2004.

ISO 19005-1 (2005). Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF 1.4 (PDF/A-1). Recommendation ISO 19005-1:2005.

ISO 19005-2 (2011). Document management – Electronic document file format for long-term preservation – Part 2: Use of ISO 32000-1 (PDF/A-2). Recommendation ISO 19005-2:2011.

ISO 19005-3 (2012). Document management – Electronic document file format for long-term preservation – Part 3: Use of ISO 32000-1 with support for embedded files (PDF/A-3). Recommendation ISO 19005-3:2012.

ISO 32000-1 (2008). Document management – Portable document format – Part 1: PDF 1.7. Recommendation ISO 32000-1:2008.

ISO 8601 (2004). Data elements and interchange formats – Information interchange – Representation of dates and times. Recommendation ISO 8601:2004.

ISO/IEC 15444 (2004). Information technology – JPEG 2000 image coding system: Core coding system. Recommendation ISO/IEC 15444-1:2004.

Kowalczyk, S. T. (2015). Before the Repository: Defining the Preservation Threats to Research Data in the Lab. In Logasa Bogen II, P., Allard, S., Mercer, H., and Beck, M., editors, *Proc. 15th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2015)*, pages 215–222. ACM Press, New York, USA.

Lease, M. and Yilmaz, E. (2013). Crowdsourcing for Information Retrieval: Introduction to the Special Issue. *Information Retrieval*, 16(2):91–100.

Lemmens, B. (2014). Challenge Brief. PREFORMA PCP Project, EU 7FP, Contract N. 619568. http://www.digitalmeetsculture.net/wp-content/uploads/2014/06/PREFORMA_Challenge-Brief_v1.0.pdf.

Lemmens, B., Elfner, P., Lundell, B., Prandoni, C., and Fresa, A. (2014). Deliverable D2.2 – Tender Specifications. PREFORMA PCP Project, EU 7FP, Contract N. 619568. http://www.digitalmeetsculture.net/wp-content/uploads/2014/05/PREFORMA_D2.2_Tender-Specifications_v2.1.pdf.

Ross, S. (2012). Digital Preservation, Archival Science and Methodological Foundations for Digital Libraries. *New Review of Information Networking*, 17(1):43–68.

Sanderson, M. (2010). Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends in Information Retrieval (FnTIR)*, 4(4):247–375.

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys (CSUR)*, 34(1):1–47.

Smucker, M. D., Kazai, G., and Lease, M. (2013). Overview of the TREC 2012 Crowdsourcing Track. In Voorhees, E. M. and Buckland, L. P., editors, *The Twenty-First Text REtrieval Conference Proceedings (TREC 2012)*. National Institute of Standards and Technology (NIST), Special Publication 500-298, Washington, USA.

Soboroff, I., Nicholas, C., and Cahan, P. (2001). Ranking Retrieval Systems without Relevance Judgments. In Kraft, D. H., Croft, W. B., Harper, D. J., and Zobel, J., editors, *Proc. 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001)*, pages 66–73. ACM Press, New York, USA.

Sokolova, M. and Lapalme, G. (2009). A Systematic Analysis of Performance Measures for Classification Tasks. *Information Processing & Management*, 45(4):427–437.

Voorhees, E. M. (2000). Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management*, 36(5):697–716.