Evaluation of Conformance Checkers for Long-Term Preservation of Multimedia Documents

Nicola Ferro
University of Padua, Italy
ferro@dei.unipd.it

Gianmaria Silvello
University of Padua, Italy
silvello@dei.unipd.it

Erik Buelinckx
Royal Institute for Cultural Heritage
KIK-IRPA, Belgium
erik.buelinckx@kikirpa.be

Boris Doubrov
Dual Lab, Belgium
boris.doubrov@duallab.com

Antonella Fresa
Promoter, Italy
fresa@promoter.it

Magnus Geber
Swedish National Archives
Riksarkivet, Sweden
magnus.geber@riksarkivet.se

Klas Jadeglans
Swedish National Archives
Riksarkivet, Sweden
klas.jadeglans@riksarkivet.se

Börje Justrell
Swedish National Archives
Riksarkivet, Sweden
borje.justrell@riksarkivet.se

Bert Lemmens
Center of Expertise in Digital
Heritage, PACKED, Belgium
bert@packed.be

Jerôme Martinez
MediaArea, France
jerome@mediaarea.net

Víctor Muñoz
Easy Innova, Spain
victormunoz@easyinnova.com

Sònia Oliveras
Records Management, Archives and
Publications Service
Girona City Council, Spain
soliveras@ajgirona.cat

Claudio Prandoni
Aedeka, Italy
prandoni@aedeka.com

Dave Rice
MediaArea, France
dave@dericed.com

Stefan Rohde-Enslin
Institute for Museum Research
Prussian Cultural Heritage
Foundation, SPK, Germany
s.rohde-enslin@smb.spk-berlin.de

Xavi Tarrés
Easy Innova, Spain
xavitarres@easyinnova.com

Erwin Verbruggen
Sound and Vision, The Netherlands
everbruggen@beeldengeluid.nl

Benjamin Yousefi
Swedish National Archives
Riksarkivet, Sweden
Benjamin.Yousefi@riksarkivet.se

Carl Wilson
Open Preservation Foundation, UK
carl@openpreservation.org

ABSTRACT

We develop an evaluation framework for the validation of conformance checkers for the long-term preservation. The framework assesses the correctness, usability, and usefulness of the tools for three media types: PDF/A (text), TIFF (image), and Matroska (audio/video). Finally, we report the results of the validation of these conformance checkers using the proposed framework.

CCS CONCEPTS

• **Information systems → Digital libraries and archives**;

KEYWORDS

long-term preservation, conformance checking, evaluation

ACM Reference Format:
Nicola Ferro, Gianmaria Silvello, Erik Buelinckx, Boris Doubrov, Antonella Fresa, Magnus Geber, Klas Jadeglans, Börje Justrell, Bert Lemmens, Jerôme Martinez, Víctor Muñoz, Sònia Oliveras, Claudio Prandoni, Dave Rice, Stefan Rohde-Enslin, Xavi Tarrés, Erwin Verbruggen, Benjamin Yousefi, and Carl

JCDL '18, June 3–7, 2018, Fort Worth, TX, USA
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5178-2/18/06…$15.00
https://doi.org/10.1145/3197026.3197037

Wilson. 2018. Evaluation of Conformance Checkers for Long-Term Preservation of Multimedia Documents. In *JCDL ’18: The 18th ACM/IEEE Joint Conference on Digital Libraries, June 3–7, 2018, Fort Worth, TX, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3197026.3197037

1 INTRODUCTION

“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” [29, p. 45]. Since preservation aims at capturing the very essence of digital objects it is often associated with life cycles [28], preservation actions, and overall preservation frameworks and there is often the need to evaluate them and choose among them [3, 4, 20].

Memory institutions, in Europe and elsewhere, are facing a situation when transfers of electronic documents or other electronic media content for long term preservation are continuously increasing. Data are normally stored in specific file formats for documents, images, sound, video etc. that are produced by software from third-party providers controlled neither by the institution that produces the files, nor by the one holding the archive. There is a risk that data objects meant for preservation, passing through an uncontrolled generative process, can put at risk the whole preservation exercise.

PREFORMA[1] is an EU co-funded *Pre-Commercial Procurement (PCP)* project, whose main objective is to give memory institutions full control of the process for testing the conformity of files to be ingested into their archives for long-term preservation.

PCP operates by clustering together stakeholders in a given domain – memory institutions in our case – which group together in order to face a common technological challenge. The stakeholders consortium is in charge of describing the expected technological solution, specifying its needed features and characteristics, and defining how alternative approaches will be compared and assessed in order to understand their pros and cons. The consortium is responsible for monitoring the progress of the suppliers work towards the first product testing and for evaluating the final solution developed by the suppliers in order to understand which one best fits with their actual needs.

The main objective of PREFORMA is the development and deployment of an open source software licensed reference implementation for file format standards aimed for any memory institution (or other organisation with a preservation task) wishing to check conformance with a specific standard. This reference implementation, called the *conformance checker* consists of a set of modular tools which, in this paper, we validate against specific implementations of specifications of standards relevant to memory institutions for preserving their different kind of data objects.

A conformance checker:

- verifies whether a file has been produced according to the specifications of a standard file format, and hence,
- verifies whether a file matches the acceptance criteria for long-term preservation by the memory institution,
- reports in human and machine readable format which properties deviate from the standard specification and acceptance criteria, and
- performs automated fixes for simple deviations in the metadata of the preservation file.

The conformance checker software developed by PREFORMA is intended for use within the *Open Archival Information System (OAIS)* Reference Framework [23] and development is guided by the user requirements provided by the memory institutions that are part of the PREFORMA consortium. The conformance checker facilitates memory institutions to obtain sufficient control of the information in an OAIS Archive, provided to the level needed to ensure long-term preservation [33]. In particular, the conformance check enables implementation of the following OAIS functions [33]: *Quality assurance* at ingestion, validating the successful transfer of the *Submission Information Package (SIP)* to the temporary storage area; *Generate AIP* at ingestion, transforming one or more SIPs into one or more *Archival Information Packages (AIPs)* that conform to the Archive’s data formatting standards and documentation standards; and, *Archival Information Update* at ingestion, providing a mechanism for updating (repackaging, transformation) the contents of the Archive.

The media types addressed by PREFORMA are: (i) *text* for establishing a reference implementation for PDF/A [24–26]; (ii) *images* for establishing a reference implementation for uncompressed TIFF [21, 22]; and, (iii) *audio-video* for establishing a reference implementation for an audiovisual preservation file, using FFV1, [2] Dirac [3] or JPEG2000 [27] for encoding video or moving image, uncompressed LPCM [19] for encoding sound and MKV [4] or OGG [5] for wrapping audio- and video-streams in one file.

This paper described the evaluation framework which has been developed for validating the PREFORMA conformance checkers. It consists of three phases: the first one is quantitative and system-oriented: the second one is qualitative and focuses on usability; and, the third one is also qualitative and targets usefulness.

The paper is organized as follows: Section 2 discusses related works; Section 3 describes our evaluation framework; Section 4 introduces the experimental collections we have developed; Section 5 presents the results of the three phases of the evaluation applied to the conformance checkers developed in PREFORMA; finally, Section 6 draws some conclusions and outlooks future work.

2 RELATED WORK

The idea of benchmarking tools for preservation is gaining more and more traction recently [7] and we share a similar approach with [9], 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;

[1] http://www.preforma-project.eu/

[2] http://www.ffmpeg.org/~michael/ffv1.html

[3] http://diracvideo.org/

[4] http://www.matroska.org/

[5] https://xiph.org/ogg/

- *performance measures* are qualitative or quantitative measurements taken by a human or a machine to calculate how fit the subject is for the task.

Creating a benchmark for complex tasks as long-term preservation, is expensive and requires to overcome multiple issues as highlighted by [10] for evaluating text extraction tools; they pointed out that the lack of proper datasets and related ground truth is the main obstacle for evaluating these tools. The construction of an experimental collection for these tasks is difficult because of the complex input space and the costs related to ground truth creation. Hence, [10] opted for developing a completely synthetic dataset, whereas we develop an experimental collection which follows the main principles of the Cranfield framework as [10] does, but it is composed of both real and synthetic files. Moreover, we extend this framework by adding two more phases to better assess usability and usefulness aspects of the evaluated tools.

In [6] the main scope of the PREFORMA project was described along with the challenges that it has to overcame to carry out the selection of three conformance checkers for text, images and audio/video. In [12], the selection process and methodology has been described as well as the main ideas for evaluating the selected tools; in particular, the evaluation as a classification task was defined in that paper, but there are no details about the overall evaluation framework composed of the three phases and the experimental results we describe and discuss in this paper.

3 EVALUATION FRAMEWORK

The framework we adopted to evaluate the conformance checkers is composed of three phases as summarized in Figure 1; each phase evaluates a different aspect of the conformance checkers: *correctness*, *usability* and *usefulness*. This framework is inspired by the triptych model [17] proposed to evaluate digital libraries where performance, usability and usefulness are evaluated holistically by considering systems, content and users. This framework can be also a further means to augment the reproducibility of the evaluated tools [13], the possibility to cite the data produced by the tools and by the evaluation process [18] and their overall quality in the broader context of the digital library quality model [11, 15].

The evaluation framework is built around a close collaboration between technical and domain experts and each evaluation phase instructs, in a continuous feedback fashion, the developers which improve the tools and address the detected issues.

3.1 Phase 1: Correctness

The first phase evaluates the correctness of the conformance checkers and it is aimed at understanding if these softwares respect the respective standards for long-term preservation, as described in [12] we briefly summarize here. Hence, the focus is on the systems and we relied on a Cranfield-like methodology [8]. The Cranfield paradigm makes use of experimental collections composed of a collection of documents of interest, a set of topics and a ground-truth which, given a document in the collection and a topic, determines the relevance of the document with respect to the considered topic.

We instantiate the Cranfield-based evaluation as a classification task, where we consider a set of classes, say C, as topics and the ground-truth is given by the correct labels assigned to each document for a given class [30]. Since the goal of the conformance checkers is to validate documents against their respective standards, we can determine, for each document, whether it is compliant, it suffers from issue 1, issue 2, and so on. Therefore, documents are labeled 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.

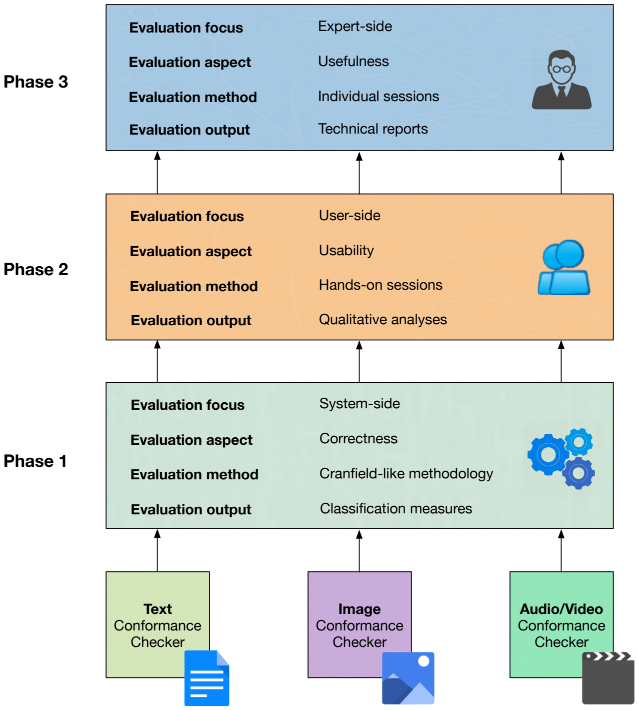

Figure 1: The evaluation framework capturing three aspects of the conformance checkers: correctness, usability and usefulness.

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.

In terms of the approach proposed by [9], we have that: the *motivating comparison* is given by the need of assessing conformance checkers; the *task sample* is defined by the identified classes $C_i \in C$, the gathered documents, and the ground-truth.

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. Documents must be representative of the different classes C_i against which we need to evaluate conformance checkers. In particular, we cannot have empty classes, i.e. classes without documents in the experimental collection, and the cardinality of each class, i.e. the number of documents 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 by the cardinality of the classes, in order to confront conformance checkers with realistic settings. Secondly, we should not introduce any bias in the evaluation measurement and process due to an uncontrolled and excessive discrepancy in the cardinality of the classes.

Ground-truth creation is a 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 [31] to crowdsourcing [1]. Unfortunately, in the context of reference, 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 issues need to be considered during ground-truth creation. The first one 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 inter-assessor agreement and see whether on this highly specialised task it will have similar ratios as those for ad-hoc retrieval [34], i.e. in the range 30%–50%, or whether discrepancies from previously known tasks will arise.

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. Indeed, we quantified 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 [32]. In our context each class C_i represents a possible mis-conformance with respect to a reference standard with the sole exception of the class C_0 which represents documents fully conforming to the standard. Thus, we define as *True Positve (TP)* the set of documents that a conformance checker has correctly labeled as belonging to class C_i; *True Negative (TN)* as the set of documents correctly labeled as not belonging to class C_i; *False Positive (FP)* as the set of documents incorrectly labeled as belonging to class C_i; and, *False Negative (FN)* as the set of documents incorrectly labeled as not belonging to class C_i.

Note that the meaning of the confusion matrix changes when considering C_0 – the class of compliant documents – 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.

Therefore, we rely on two main evaluation measures able both to give a general account of conformance checkers performances and to deal with this duality between FNs and FPs: *Accuracy* and *Area under the curve (AUC)*.

Accuracy measures the overall effectiveness [32] of a conformance checker as

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

AUC measures the ability of a conformance checker to avoid false classification [32] as

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

In order to obtain a single score for each conformance checker across all the categories C_i, we use a *macro-averaging* approach [30], which computes the arithmetic mean of the above measures over all the categories C_i.

Moreover, 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. $C_0 \cap C_i = \emptyset \, \forall i, \ i \neq 0$.

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

$$\text{Consistency}_i = 1 - \frac{|C_0 \cap C_i|}{|C_i|} \tag{3}$$

where N is the total number of classes, excluded C_0. Note that consistency is different from the evaluation measures typically used in classification or clustering and serves the specific purpose of assessing the degree of separation between the compliant and not-compliant classes. This measure has been extensively used as a tool for refining the experimental collection while building it as we discuss below.

3.2 Phase 2: Usability

The second phase evaluates the usability of the conformance checkers. These tools are thought for memory institutions where domain experts have to decide if a given document – i.e. a text document, an image or an audio/video file – is compliant with the current standards for long-term preservation. To this end, usability is evaluated by means of user studies/focus groups with domain experts from different memory institutions.

The user studies have the form of hands-on sessions where the conformance checkers are described to the domain experts in order to get them to know the basic functionalities and how they work. In a second phase, the experts are divided into groups, usually one per conformance checker, and are asked to download, install and use the softwares by testing some files from their institutions. The files to be tested are selected by the experts themselves and so the hands-on session organizers have no control over the files to be tested; this guarantees that the user studies are unbiased and as close as possible to a real-world work scenario.

The usability of the conformance checkers is evaluated by assessing specific functionalities of the tools such as installation complexity, user interface, interpretation of the results, documentation and difference between desktop and web-based version of the checkers. Furthermore, after the use on the tools, other aspects such as the potential for innovation, ease of use, scalability, extensibility and interoperability are assessed by the domain experts.

The participants to the user studies are asked to compile questionnaires composed of several questions to be answered with a number from 1 (very poor) and 5 (very good) and other questions requiring a short textual answer.

3.3 Phase 3: Usefulness

The third phase is carried out by an evaluation committee composed of selected domain and technical experts that assess the final version of the conformance checkers. Moreover, the developers of the conformance checkers produce a final report where the results of the evaluation and all the fixes and improvements required by the experts are described.

The domain experts collect the files provided by some selected external institutions and validate them with the right conformance checker. Then, they verify if any issue arise and assess the XML and HTML report produced by the tool. Finally, they write a report describing the weaknesses of the tested tool, which is then used by the developers to improve their product. The technical experts verify if the tool compiles and can be used with different operating systems, if the installation process prompts any problems as well as if the validation process works correctly. At the end, they also produce a report describing possible problems to be fixed.

The report written by the developers comprises considerations about the files used for testing, the improvements made during the evaluation phases, the innovative aspects of the tools with respect to other existing tools, standardization efforts, dissemination activities, the partaken open-source approach, future plans for improvements and additional information, if any.

This evaluation phase allows us to further review the conformance checkers, to improve and fix some remaining issues and to produce a thorough documentation to be used as reference for future developments of the tools also by third-party agents.

4 EXPERIMENTAL COLLECTIONS

We created three shared and publicly available experimental collections, one for each media type targeted by PREFORMA, based on the Cranfield framework – i.e. a corpus of documents, a set of test classes of conformance and a ground truth created by experts.

The experimental collections we defined have a two-fold goal: *training*: the collections are used as a development tool to check the conformance checker functionalities while they are under construction in order to get immediate feedback and correct the issues that may arise; *testing*: the collections are used to evaluate the correctness of the checkers by a third-party organization not involved in their development, thus providing an unbiased assessment of the tools.

4.1 Corpora

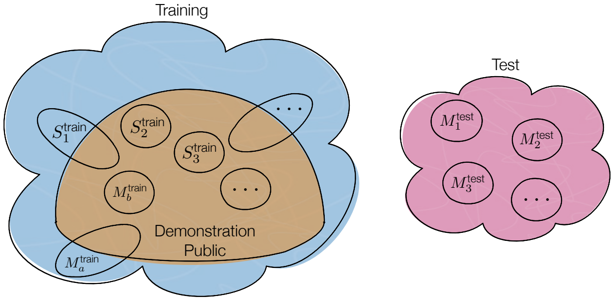

Figure 2: The training and test document corpora.

Figure 2 shows the structure of the document corpora we defined and made available for the development and the evaluation of the conformance checkers. The main distinction is between:

- *training dataset*: aimed at driving and facilitating the design and development of the conformance checkers, as well as show casing their functionalities.
- *test dataset*: aimed at evaluating and testing the tools.

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

Both training and test dataset are associated with ground-truth specifying the correct labels for the documents in the dataset but the ground-truth associated with the test data 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 providing the documents. The training dataset is constituted by two parts: a *demonstration* one, which can be used to show casing the tools; a *private* part, which is used internally by each developer for designing, developing, and testing its own system.

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 problem or critical issue for a given preservation format, as proposed also by [2]. The latter are data actually managed by memory institutions for their preservation duties. Both the training and the test datasets comprise both synthetic and real data.

4.2 Classes and Ground-Truth

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:

- one 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;
- two experts from memory institutions;
- one expert from developers.

For each class it is specified a unique identifier, a short name and a brief description. The class definition process was two-fold. Firstly, the experts defined the classes on the basis of the standard and in a second moment they revised them by removing the classes for which there were not enough files belonging to them, were not correctly specified or were too generic. Here we report a few examples of classes for each media type [14] – TC is a for text classes, IC is for image classes, and AV is for audio/video classes:

- 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
- 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
- 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
- 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 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 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 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 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 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]

The experimental collection for text initially contained 88 classes that after a refinement were reduced to 78; 10 classes have been removed. The remaining classes verify if the PDF/A standard is met by the test files; for instance, they check if they contain annotations (e.g. sound, movies, 3D objects), use any denied compression (e.g. LZW), contain interpolated images, content layers, transparencies, executable scripts, action forms, javascript and so on.

The experimental collection of images initially contained 44 classes of which only one was removed after the refinement process.Among other things, the other classes verify the size of the TIFF offset, tag type, channels error, dimensions error, resolution error, TIFF signature, byte order, photometric interpretation, color tags, lossy compression and size of the uncompressed TIFF-file.

Finally, the experimental collection of audio/video initially had 69 classes, reduced to 56 after refinement that removed generic classes and classes targeting deprecated features of the Matroska standard (still under revision at the time of writing). Classes target files element data size length, element size byte length limit, non-Ascii data in string, missing header, coder type and so on.

4.3 Resources

All the resources required to reproduce the results presented in this paper as well as the source code of the tested conformance checkers are publicly available. The experimental collections for text, image and audio/video for both training and testing are available in the PREFORMA GitHub repository: http://github.com/preforma.

The source code of the conformance checkers is available at the following URLs:

- text conformance checker (veraPDF) source code: http://github.com/verapdf
- image conformance checker (DPF Manager) source code: http://github.com/EasyinnovaSL/DPFManager
- audio/video conformance checker (MediaConch) source code: http://github.com/MediaArea/MediaConch_SourceCode/

The evaluation was conducted by using the open source MATTERS library: http://matters.dei.unipd.it/.

5 EXPERIMENTAL RESULTS

5.1 Phase 1: Correctness

The Cranfiled-based evaluation framework has been iterated several times with the purpose of spotting issues and improving the tools until the performances reached a high level of correctness.

The tested tool for the text media type was the text conformance checker developed by VeraPDF. [6] In Table 1 we can see the values of accuracy, AUC and consistency averaged over all the text classes. The results are pretty good since accuracy is very close to the maximum value, whereas AUC could be further improved [16].

As we can see the text conformance checker reports some misclassification for the files within the class `TC000` which is the class with "correct files"; this means that the checker returns some false negative that would need a further check before being accepted for long-term preservation by a memory institution. The classes which

[6] http://verapdf.org/home/

Table 1: Average values of accuracy, AUC and consistency for text, image and audio/video media types. The minimum value is 0 and the maximum is 1.

Media type	Accuracy	AUC	Consistency
Text	0.9812	0.8495	1.0000
Image	1.0000	1.0000	1.0000
Audio/Video	0.9977	0.9621	1.0000

contain more mis-classified files are `TC010` (image encoded interpolation), `TC016` (has attachment of any kind of resource), `TC017` (document non-PDF/A attachment), `TC040` (has spacings around keywords 'obj', 'endobj', 'stream', 'end- stream', 'xref') and `TC064` (does not have functionality required for Conformance Level U).

For images the tested conformance checker was DPF manager TIFF conformance checker. [7] As we can see from Table 1, this checker gets 100% accuracy, AUC and consisntecy for all the classes.

For audio/video the tested conformance checker was MediaConch [8] for Matroska, FFV1 and PCM. From Table 1, we see that this chacker behaves pretty well on average; the two classes failing to obtain 100% accuracy are `AV001` (The first Element ID must equal `0x172351395` (EBML Header) [EBML/EBML-ELEM-START]) and `AV061` (MKV is not at least version 4 [Matroska/MKV-V4+]).

5.2 Phase 2: Usability

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

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

5.2.1 User study. We organized two hands-on sessions with a total of 21 participants that tested the tools; fourteen participants were interested in the text format, sixteen in the images and thirteen in the audio/video format (of course there were participants interested in more than one media type). The participants were asked:

- "Is there any other file format for which you would be interested in having a conformance checker, besides those already covered in PREFORMA (PDF/A, TIFF, Matroska/FFv1)?"
  The vast majority answered "No", three wanted also the TXT format to be covered and one wanted also the JPEG format to be covered.
- How would you envisage to use the results of the open source projects in your legacy environment and/or in your digital archiving and preservation initiatives?
  Three answered "Web", two (from the cultural heritage domain) "Integration to legacy tools", eight "Integration in legacy system" and six "Standalone service via web".
- Do you expect that the PREFORMA's results will have a positive impact on the workflow of your institution/organisation? Seven answered "yes" because "It could improve the workflow, and relations with other institutions and agencies", "To test the tools and to find a possible integration with our system", "to increase the awareness of the issue" and "to increase discussion about policies and standardisation"; one answered "no" (i.e. the electronic adm. consultant) and all the others answered "too early to say".

All the participants tested all the conformance checker and we asked how they would rate their experience in using the tool(s) in terms of performance, usability, potential and other related aspects. The opinion were expressed with a number from 1 (very poor) to 5 (very good) and with a free text. The answers are reported in Table 2. We see that the user evaluation is quite good for all the considered aspects, even though they had difficulties at assessing the quality of documentation, scalability, interoperability and licenses. These features are hard to assess without enough time, long use of the tools and the required technical expertise; these aspects were further targeted in phase 3 of the evaluation.

Another question was: "How would you rate your experience in using the tool(s) in terms of specific functionality?" The answers are reported in Table 3 (one person did not answer).

To the question "What's the one single thing that you learnt today?", the most common answers were:

- The importance of some details, too often forgotten.
- Creating policies.
- How to create policy for audio/video.
- There is a good and interesting tool which deserves to be analysed and tested deeper.
- You can implement your own checking rules.
- There is a community working to answer my doubts.
- Application of correct parameters for the long-term preservation.
- More consciousness of the obsolescence.
- The digital preservation could be possible because we knew the right tools.
- Identifying and fix file formats and know a new software to do that.
- There is a software that can help to the administrations to know better their TIFF files for digital preservation.
- Using DPF Manager to check TIFF files.

5.2.2 Focus groups. In addition to the user study presented above, also two focus groups were organized. The first one was focused on the text media format and it was composed of 21 people:

- 12 came from state agencies and organisations, including 2 from public museums and archives and 3 from the offices of the Swedish Parliament and of the Swedish Government;

[7] http://dpfmanager.org/
[8] https://mediaarea.net/MediaConch/

Table 2: How would you rate your experience in using the tool(s) in terms of performance, usability, potential and other related aspects? From 1 (very poor) to 5 (very good) with a 0.5 step.

	Innovation potential	Performance	Usability	Documentation	Scalability	Interoperability	Licenses
User 01	4	4	4	4	-	-	4
User 02	4	4	3	3	4	3	-
User 03	4	4	5	4	4	-	-
User 04	4	5	5	5	5	5	5
User 05	-	-	5	5	-	-	-
User 06	4	5	5	4	4	4	4
User 07	5	5	4	4	4	4	4
User 08	5	4	4	4	-	-	-
User 09	4	5	5	5	5	5	5
User 10	-	-	5	5	-	-	-
User 11	4	4	4	4	-	-	-
User 12	5	5	4	-	-	-	-
User 13	5	4	4	2	4	4	5
User 14	4	2	4	2	4	4	4
User 15	5	2.5	3	-	-	-	5
User 16	3	2	2	2	3	-	-
User 17	4	3	4	3	-	-	-
User 18	4	5	4	-	-	2	-
User 19	4	4	3	-	-	-	4
User 20	2.5	3	2	-	-	-	-
User 21	2.5	3	4	-	-	4	-
Avg.	4.05	3.86	3.95	3.73	4.11	3.88	4.44
Std. Dev.	0.76	1.03	0.92	1.10	0.60	0.93	0.53

Table 3: How would you rate your experience in using the tool(s) in terms of specific functionality? From 1 (very poor) to 5 (very good) with a 0.5 step.

	Installation and configuration	User Interface	Command Line Interface	Conformance checking	Policy creation and checking	Results interpretation	Metadata fixing	Documentation	Web-based version
User 01	3	5	-	5	4	4	-	3	-
User 02	5	4	-	-	5	-	5	-	-
User 03	3.5	4.5	-	5	5	4	-	4	-
User 04	5	5	5	5	5	4	4	-	-
User 05	5	4	-	-	5	4	-	4	-
User 06	4	-	-	-	-	-	-	-	-
User 07	1	4	4	4	3	3	-	-	-
User 08	2	4	4	3	2	4	3	2	4
User 09	-	-	-	-	-	-	-	-	-
User 10	4	2	2	2	1	3	-	-	-
User 11	3	4	4	4	4	4	-	-	-
User 12	3.5	4.5	-	5	5	5	-	-	5
User 13	2	3	3	4	2	4	-	-	-
User 14	3.5	3	-	2	2	2.5	-	-	-
User 15	4	4	-	4	5	3	-	-	-
User 16	4	4	4	4	4	4	4	4	4
User 17	4	4	4	5	4	3	5	3	3
User 18	5	5	4	5	4	5	5	4	5
User 19	5	5	5	5	5	5	5	5	5
User 20	5	4	4	-	5	5	-	5	5
Avg.	3.76	4.06	3.91	4.13	3.89	3.91	4.43	3.78	4.43
Std. Dev.	1.71	0.78	0.83	1.06	1.32	0.80	0.79	1.00	0.80

- 5 came from communal agencies, 3 from local and 2 from regional ones:
- 3 came from small-medium enterprises targeting IT and archiving;
- 1 came from a private archival institution.

The focus group started with a presentation about what is conformance checking and what we mean with policy checking. Afterwards, there was a brief overview of the PDF/A format. All participants were to some extent familiar with the format, so this was more of an update and to equalize their knowledge.

All participants had to bring with them their own laptops (with Java and veraPDF version 1.4.6 installed) and examples of PDF/A files from their organisations. It was explained to the participants: how to read reports and failure messages, how to identify and interpret failures, possible actions to solve failures, how to formulate a policy and how to test a policy.

Each item was explained by means of a demonstration and was followed by practical exercises, where the participants used their own files or test files downloaded from Internet.

The focus group ended with a discussion to evaluate the day. The expectations expressed in the beginning of the seminar were mainly focusing on learning more about PDF/A in order to better understand its "pros and cons" but also to learn about validation and the text conformance checker. The feedback from the participants was very positive. Some commented that conformance checking requires knowledge about the format that most curators of digital object do not have today. More "fixers" was also asked for; the metadata fixer could be complemented with more simple fixing.

The second focus group was focused on the the image media type and composed of 22 participants mainly from small/medium-sized museums comprising directors, curators and some IT-people. At the beginning, there was a general introduction into file-formats for digital preservation. Four participants brought a laptop and some TIFF-files along to be tested. There was no problem in downloading and installing the software, but many of the participants were convinced that the IT-people responsible for their museum would have troubles installing the software. This impression was not supported by any evidence, but it has to be taken into account in order to simplify as much as possible the download and installation of the tools.

After the installation of the image conformance checker (DPF-Manager), the functionalities available in the windows version were explained; the participants were able to understand "conformance", "policy", etc. Most functionalities worked smoothly. Some functioning issues came up when using MS Windows as operating system.

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

At the end the participants agreed that the DPF-Manager is a valuable tool for their digitization-work, not only for digital preservation but also to be used when checking image-files produced by external companies in the framework of a digitization project of the museum. There was room for improvement: The messages should be in multilingual (German in this case) or at least there should be the possibility to set up languages other than English. It might be good to have some hints on software tools to correct found errors.

In general, it took some time to make clear, that conformance to baseline-TIFF is necessary for digital preservation. But after this was clarified, participants were eager to use the DPF-Manager tool.

5.3 Phase 3: Usefulness

The issues highlighted by the user study and the focus groups were analysed by the developers that improved the conformance checker tools accordingly. In particular, the documentation of the tools has been improved and the download and installation procedures have been simplified. Afterwards, the tools have been evaluated by some technical and domain experts that provided three final reports (one per media type) with the overall evaluation and additional suggestions for improving the tools.

5.3.1 Text Media Type. The domain expert underlined that it was a beneficial experience to test the veraPDF software since many national libraries are currently looking to develop digital archives which will need to have tools to validate their content. After an internal testing phase with some PDF/A files, no issues have been detected and the output of the conformance checker was consistent with the one of another checker. The tool works as expected, produces XML and HTML reports and saves them as it should. Nevertheless, the reports could be more user friendly for non-technical users.

The technical expert highlighted no problems with the download, compilation, installation and use of the conformance checker. Moreover, the expert highlighted that some previous issues have been resolved and documented by the developer.

5.3.2 Image Media Type. The domain and technical experts highlighted that this tool has been tested with the intention of using it in a high-volume automated productions system, hence the GUI is less important for them with respect to the command line. The DPF Manager worked well and fast enough and the experts thought that it should be made part of digital preservation workflows in museums. But, they discovered that it is also a useful tool for digitization projects. Often digitization in museums is done by external companies; hence, with DPF Manager the museums are enabled to check the quality of the work of these companies.

The domain expert wrote that he was very happy with the program and with the fast reaction when asking for some updates (e.g., to be able to check a certain additional policy-feature, the addition of basic identification metadata and the statistic module).

The technical expert tried out DPF Manager for both Windows and Linux Ubuntu platforms. Nevertheless, the instructions for installing in non-GUI-Linux systems were still missing in the documentation. Overall, the experts underlined that the software (both GUI and command line) should be made easier and more intuitive because even they, as expert users, find it hard to understand how to use the checker in a more detailed or specialized way, especially when they wanted to ignore errors in selected tags.

A future development plan should account for an interface to provide easily translations of massages and reports in languages other than English.

5.3.3 Audio/Video Media Type. The domain expert stated that the main issue with MediaConch is that, to him, it is a tool for professionals. To work well for users that do not have a good technical knowledge it will have to be completed with, perhaps a knowledge base and a guide (guidance) on how to interpret conformance errors. As for now the tool provides an error with a technical detail but no guidance on how this error could be interpreted or corrected. The expert pointed out that the software should benefit from adding a knowledge base to the software or the platform on Internet. The knowledge base should cover references to common errors and guidance on how to correct common errors. It could also include hints about how to deal with a more complex type of errors. This should probably make the software available to a broader public and increase its user base. In addition, an issue with the licensing has been covered within the evaluation of the final report.

The technical expert pointed out that that the standardization process for audio/video is ongoing and this fact created problems to the developers. As long as the standardization is ongoing, the conformance checking of the standard is no serious matter. However, the other areas of checking, such as the profiles (as presets or individually defined), still make sense. Overall, the conformance checker has been tested and it works quite well.

6 CONCLUSIONS AND FUTURE WORK

This paper discussed the overall evaluation framework we developed, in the context of the PREFORMA EU project, to validate the conformance checkers for the long-term preservation of PDF/A (text), TIFF (image), and Matroska (audio/video) files. The framework assesses three aspects of the conformance checkers: (i) correctness, in a quantitative system-oriented way; (ii) usability, in a qualitative user-oriented way; and, (iii) usefulness, in a qualitative domain and technical expert oriented way. The developed framework, besides its application in the PREFORMA project context, is a general tool which can be used for evaluating also other preservation related tasks, provided that the definition of the classes and the ground-truth are tailored to the case at hand.

The paper described the three experimental collections, one for each media type, we developed and which are open source available, the results of the conformance checkers evaluation using these collections, and the outcomes for usability and usefulness gathered with user hands-on sessions and domain experts focus groups. Overall, all these types of validation indicate that the PREFORMA conformance checkers are ready to be shared open source with the memory institution community at large.

Future work will concern the refinement of the framework by introducing a notion of severity for the classes to be used to weight the impact of each mis-conformance and to obtain a more fine-grained validation of the tools. Furthermore, we plan to extend the defined classes with a new set based on policies and best practices of the memory institutions and not only on strict conformance to the standards.

REFERENCES

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

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

[3] C. Becker, K. Duretec, and A. Rauber. The Challenge of Test Data Quality in Data Processing. *ACM Journal of Data and Information Quality (JDIQ)*, 8(2), 2016.

[4] C. Becker and A. Rauber. 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, 2011.

[5] D. Calvanese, D. De Nart, and C. Tasso, editors. *Digital Libraries on the Move – Proc. 11th Italian Research Conference on Digital Libraries (IRCDL 2015)*. Communications in Computer and Information Science (CCIS) 612, Springer, Heidelberg, Germany, 2016.

[6] L. Cappellato, N. Ferro, A. Fresa, M. Geber, B. Justrel, B. Lemmen, C. Prandoni, and G. Silvello. The PREFORMA Project: Federating Memory Institutions for Better Compliance of Preservation Formats. In Calvanese et al. [5], pages 86–91.

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

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

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

[10] K. Duretec, A. Rauber, and C. Becker. A Text Extraction Software Benchmark Based on a Synthesized Dataset. In *2017 ACM/IEEE Joint Conference on Digital Libraries, JCDL 2017*, pages 109–118. IEEE Computer Society, 2017.

[11] N. Ferro. Quality and Interoperability: The Quest for the Optimal Balance. In I. Iglezakis, T.-E. Synodinou, and S. Kapidakis, editors, *E-Publishing and Digital Libraries: Legal and Organizational Issues*, pages 48–68. IGI Global, USA, 2010.

[12] N. Ferro. Proposal for an Evaluation Framework for Compliance Checkers for Long-term Digital Preservation. In M. Agosti, M. Bertini, S. Ferilli, S. Marinai, and N. Orio, editors, *Digital Libraries and Multimedia Archives – Proc. 12th Italian Research Conference on Digital Libraries (IRCDL 2016)*, pages 125–136. Communications in Computer and Information Science (CCIS) 701, Springer, Heidelberg, Germany, 2016.

[13] N. Ferro. Reproducibility Challenges in Information Retrieval Evaluation. *ACM Journal of Data and Information Quality (JDIQ)*, 8(2):8:1–8:4, January 2017.

[14] N. Ferro, E. Buelinckx, B. Doubrov, K. Jadeglans, B. Lemmens, J. Martinez, V. Muñoz, C. Prandoni, D. Rice, S. Rohde-Enslin, X. Tarrés, E. Verbruggen, B. Yousefi, and C. Wilson. Deliverable D8.1R2 – Competitive Evaluation Strategy. PREFORMA PCP Project, EU 7FP, Contract N. 619568, October 2016.

[15] N. Ferro and G. Silvello. Towards a Semantic Web Enabled Representation of DL Foundational Models: The Quality Domain Example. In Calvanese et al. [5], pages 24–35.

[16] N. Ferro, G. Silvello, E. Buelinckx, B. Doubrov, M. Geber, K. Jadeglans, J. Martinez, V. Muñoz, D. Rice, S. Rohde-Enslin, X. Tarrés, E. Verbruggen, B. Yousefi, and C. Wilson. Deliverable D8.6 – Testing Report. PREFORMA PCP Project, EU 7FP, Contract N. 619568, October 2017.

[17] N. Fuhr, G. Tsakonas, T. Aalberg, M. Agosti, P. Hansen, S. Kapidakis, C.-P. Klas, L. Kovács, M. Landoni, A. Micsik, C. Papatheodorou, C. Peters, and I. Sølvberg. Evaluation of Digital Libraries. *International Journal on Digital Libraries*, 8(1):21–38, 2007.

[18] Silvello G. Theory and practice of data citation. *JASIST*, 69(1):6–20, 2018.

[19] IEC 60958. Digital audio interface - Part 1: General. Standard IEC 60958-1 Ed. 3.1 b:2014, 2014.

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

[21] ISO 12234-2. Electronic still-picture imaging – Removable memory – Part 2: TIFF/EP image data format. Recommendation ISO 12234-2:2001, 2001.

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

[23] ISO 14721. Space data and information transfer systems – Open archival information system (OAIS) – Reference model. Recom. ISO 14721:2012, 2012.

[24] ISO 19005-1. 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, 2005.

[25] ISO 19005-2. 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, 2011.

[26] ISO 19005-3. 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, 2012.

[27] ISO/IEC 15444. Information technology – JPEG 2000 image coding system: Core coding system. Recommendation ISO/IEC 15444-1:2004, 2004.

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

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

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

[31] I. Soboroff, C. Nicholas, and P. Cahan. Ranking Retrieval Systems without Relevance Judgments. In D. H. Kraft, W. B. Croft, D. J. Harper, and J. Zobel, 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, 2001.

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

[33] The Consultative Committee for Space Data Systems (CCSDS). Reference Model for an Open Archival Information System (OAIS). Magenta Book, Issue 2. Recommended Practice CCSDS 650.0-M-2, http://public.ccsds.org/publications/archive/650x0m2.pdf, June 2012.

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