Proposal for an Evaluation Framework for Compliance Checkers for Long-Term Digital Preservation

Nicola Ferro(✉)

Department of Information Engineering,
University of Padua, Padua, Italy
ferro@dei.unipd.it

Abstract. In this paper, we discuss the problem of how to model and evaluate tools that allow memory institutions to check the conformance of documents with respect to their reference standards in order to ensure their appropriateness for long-term preservation. In particular, we propose to model the conformance checking problem as a classification task and to evaluate it as a multi-classification problem using a Cranfield-like approach.

1 Introduction

The *PREservation FORMAts for culture information/e-archives (PREFORMA)*[1] project is a *Pre-Commercial Procurement (PCP)* project focused on conformity checking of ingested files for the long-term preservation [8]. The main objective of the project is the development and deployment of an open source software licensed reference implementation for file format standards aimed at 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*, will consist of a set of modular tools which will be validated against specific implementations of specifications of standards relevant to the PREFORMA project and used by the European 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.

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

© Springer International Publishing AG 2017
M. Agosti et al. (Eds.): IRCDL 2016, CCIS 701, pp. 125–136, 2017.
DOI: 10.1007/978-3-319-56300-8_12

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 media types addressed by PREFORMA are: (i) *electronic documents* 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] for encoding video or moving images, uncompressed LPCM [19] for encoding sound and MKV[3] for wrapping audio- and video-streams in one file.

Evaluation and validation of the developed conformance checkers is a primary concern in PREFORMA and this paper describes the overall approach and framework we are going to apply to assess the performances of the developed tools.

The paper is organized as follows: Section 2 presents some related works in the digital preservation area; Section 3 explains how we frame the conformance checking process as a classification task; Section 4 discusses how we evaluate the performances of the developed conformance checkers; finally, Sect. 5 draws some conclusions and presents an outlook for future work.

2 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" [29, p. 45]. Since preservation aims at capturing the very essence of digital objects it is often associated with life cycles [27], preservation actions, and overall preservation frameworks and there is often the need to evaluate them and choose among them [6,7,20].

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 [9] and we share a similar approach with [12], 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;

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

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

- *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.

3 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 [2], 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 Fig. 1.

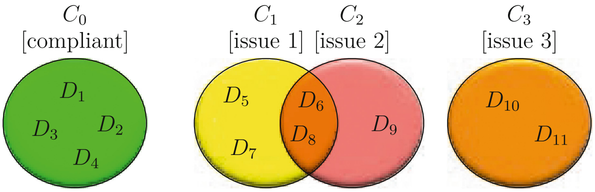


Fig. 1. Conformance checking as a classification task.

One of the challenges we have to face is how to determine the list of classes for each the media types targeted by 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 Fig. 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 Evaluating Conformance Checkers for Digital Preservation

In order to evaluate conformance checkers, we will rely on the Cranfield paradigm [10], 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 [31].

In terms of the approach proposed by [12], 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 Sect. 3, the gathered documents, as described in Sect. 4.1, and the ground-truth, as presented in Sect. 4.2; the *performance measures* are described in Sect. 4.3.

The proposed approach also enables a basic level of reproducibility [16,18] of the conducted experiments in the long-term, which we deem essential when it comes to evaluation compliance checkers for long-term preservation.

4.1 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

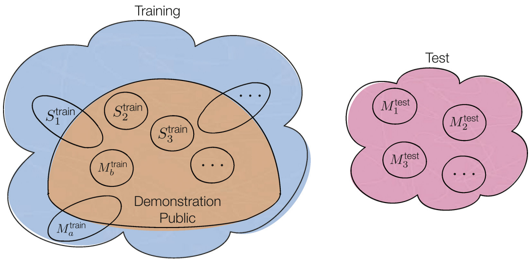


Fig. 2. PREFORMA document collections.

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 [13]. 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 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 problem or critical issue for a given preservation format, as proposed also by [5]. 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.

4.2 Ground-Truth

As it is well known [30], 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 [33] to crowdsourcing [1,28].

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.

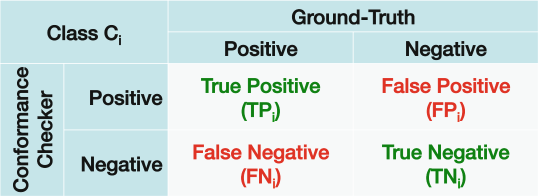
Two interesting questions will arise during ground-truth creation in PREFORMA. The first issue is that, to assess the compliance of a document, domain experts will probably also use some of the already existing tools and this may introduce circularity and bias. The second issue is to understand the problem of inter-assessor agreement and see whether on this highly specialised task it will have similar ratios as those for ad-hoc retrieval [35], 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 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 [34] shown in Fig. 3.



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)

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

Recall from Sect. 3 and Fig. 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 Fig. 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 [11], 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 [34] 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 [14,34] as

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

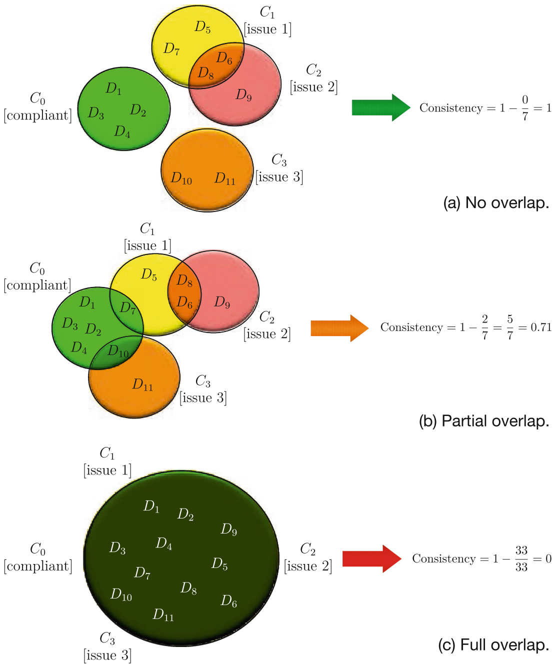
- *logistic average misclassification rate (LAM)*: is the geometric mean of the *odds* of compliance and not-compliance misclassification, converted back to a proportion [11,32]. This measure imposes no a priori relative importance on compliance and not-compliance misclassification, and rewards equally 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 $\text{logit}^{-1}(x) = \frac{e^x}{1+e^x}$.

In order to obtain a single score for each conformance checker across all the categories C_i, we will use a *macro-averaging* approach [31], which computes the arithmetic mean of the above measures over all the categories C_i.

Moreover, as explained in Sect. 3, 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



(a) No overlap.

(b) Partial overlap.

(c) Full overlap.

Fig. 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.

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 [15,31,34] or clustering [3,4] 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 $= 1$ (Fig. 4a);

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 (Fig. 4c); in the other cases, when some overlap exists, consistency is in the range $(0, 1)$ (Fig. 4b).

5 Conclusions and Future Work

In this paper we discussed how to model the process of conformance checking for long-term digital preservation and, consequently how to evaluate it. In particular, we proposed to consider conformance checking as a multi-classification problem, with the constraint that C_0, the class of compliant documents, is separated from the others. We then discussed how to instantiate the Cranfield paradigm for the specific purpose of evaluating conformance checkers, we selected the existing measures – accuracy, AUC, and LAM – that best fit this peculiar applicative context and we proposed a new measure – consistency – that assess the extent to which conformance checkers are able to keep the C_0 class separated from the other classes.

Future work will concern the application of the proposed framework in the context of the PREFORMA project, with real memory institutions, domain experts and the suppliers which are actually developing the conformance checkers for the different media types targeted by PREFORMA. In particular, we see this as an iterative process, where we will go through repeated cycles to collect larger and larger datasets, to train memory institutions and suppliers on this evaluation methodology, and to refine it. An initial account of the defined classes can be found in [17].

Acknowledgements. The reported work has been partially supported by the PREFORMA project(http://www.preforma-project.eu/), as part of the Seventh Framework Programme of the European Commission, grant agreement no. 619568.

References

1. Alonso, O.: Implementing crowdsourcing-based relevance experimentation: an industrial perspective. Inf. Retrieval **16**(2), 101–120 (2013)
2. Alpaydin, E.: Introduction to Machine Learning. The MIT Press, Cambridge (2014)
3. Amigó, E., Gonzalo, J., Artiles, J., Verdejo, M.F.: A comparison of extrinsic clustering evaluation metrics based on formal constraints. Inf. Retrieval **12**(4), 461–486 (2009)
4. Amigó, E., Gonzalo, J., Verdejo, M.F.: A general evaluation measure for document organization tasks. In: Jones, G.J.F., Sheridan, P., Kelly, D., de Rijke, M., Sakai, T. (eds.) Proceeding 36th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2013), pp. 643–652. ACM Press, New York (2013)
5. Becker, C., Duretec, K.: 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., Shipman, F. (eds.) Proceeding 13th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2013), pp. 349–358. ACM Press, New York (2013)

6. Becker, C., Duretec, K., Rauber, A.: The Challenge of Test Data Quality in Data Processing. ACM J. Data Inf. Qual. (JDIQ) **8**(2) (2016)
7. Becker, C., Rauber, A.: Decision criteria in digital preservation: what to measure and how. J. Am. Soc. Inform. Sci. Technol. (JASIST) **62**(6), 1009–1028 (2011)
8. Cappellato, L., Ferro, N., Fresa, A., Geber, M., Justrell, B., Lemmens, B., Prandoni, C., Silvello, G.: The PREFORMA project: federating memory institutions for better compliance of preservation formats. In: Calvanese, D., De Nart, D., Tasso, C. (eds.) IRCDL 2015. CCIS, vol. 612, pp. 86–91. Springer, Cham (2016). doi:10.1007/978-3-319-41938-1_10
9. Chanod, J.P., Dobreva, M., Rauber, A., Ross, S., Casarosa, V.: Issues in digital preservation: towards a new research agenda. In: Chanod, J.P., Dobreva, M., Rauber, A., Ross, S. (eds.) Report from Dagstuhl Seminar 10291: Automation in Digital Preservation, pp. 1–14. Dagstuhl Reports, Schloss Dagstuhl-Leibniz-Zentrum für Informatik, Germany (2010)
10. Cleverdon, C.W.: The Cranfield tests on index languages devices. In: Spärck Jones, K., Willett, P. (eds.) Readings in Information Retrieval, pp. 47–60. Morgan Kaufmann Publisher Inc., San Francisco (1997)
11. Cormack, G., Lynam, T.: TREC 2005 spam track overview. In: Voorhees, E.M., Buckland, L.P. (eds.) The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). National Institute of Standards and Technology (NIST), Special Publication 500–266, Washington, USA (2005)
12. Duretec, K., Kulmukhametov, A., Rauber, A., Becker, C.: Benchmarks for digital preservation tools. In: Proceeding of 11th International Conference on Preservation of Digital Objects (iPRES 2015) (2015)
13. Elfner, P., Justrell, B.: Deliverable D2.1 - Overall Roadmap. PREFORMA PCP Project, EU 7FP, Contract N. 619568, June 2014. http://www.digitalmeetsculture.net/wp-content/uploads/2014/05/PREFORMA_D2.1_Overall-Roadmap_v2.5.pdf
14. Fawcett, T.: An introduction to ROC analysis. Pattern Recogn. Lett. **27**(8), 861–874 (2006)
15. Ferri, C., Hernández-Orallo, J., Modroiu, R.: An experimental comparison of performance measures for classification. Pattern Recogn. Lett. **30**(1), 27–38 (2009)
16. Ferro, N.: Reproducibility challenges in information retrieval evaluation. ACM J. Data Inf. Qual. (JDIQ) **8**(2), 8:1–8:4 (2017)
17. Ferro, N., Buelinckx, E., Doubrov, B., Jadeglans, K., Lemmens, B., Martinez, J., Muñoz, V., Prandoni, C., Rice, D., Rohde-Enslin, S., Tarres, X., Verbruggen, E., Yousefi, B., Wilson, C.: Deliverable D8.1R2 - Competitive Evaluation Strategy. PREFORMA PCP Project, EU 7FP, Contract N. 619568, October 2016
18. Ferro, N., Fuhr, N., Järvelin, K., Kando, N., Lippold, M., Zobel, J.: Increasing reproducibility in IR: findings from the Dagstuhl Seminar on “Reproducibility of Data-Oriented Experiments in e-Science”. SIGIR Forum **50**(1), 68–82 (2016)
19. IEC 60958: Digital audio interface - Part 1: General. Standard IEC 60958–1 Ed. 3.1 b:2014 (2014)
20. Innocenti, P., Ross, S., Maceviciute, E., Wilson, T., Ludwig, J., Pempe, W.: Assessing digital preservation frameworks: the approach of the SHAMAN project. In: Spyratos, N., Kapetanios, E., Traina, A. (eds.) Proceeding of ACM International Conference on Management of Emergent Digital EcoSystems (MEDES 2009), pp. 412–416. ACM Press, New York (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. Recommendation 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. Kowalczyk, S.T.: Before the repository: defining the preservation threats to research data in the lab. In: Logasa Bogen II, P., Allard, S., Mercer, H., Beck, M. (eds.) Proceeding of 15th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2015), pp. 215–222. ACM Press, New York (2015)
28. Lease, M., Yilmaz, E.: Crowdsourcing for information retrieval: introduction to the special issue. Inf. Retrieval **16**(2), 91–100 (2013)
29. Ross, S.: Digital preservation, archival science and methodological foundations for digital libraries. New Rev. Inf. Networking **17**(1), 43–68 (2012)
30. Sanderson, M.: Test collection based evaluation of information retrieval systems. Found. Trends Inf. Retrieval (FnTIR) **4**(4), 247–375 (2010)
31. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. (CSUR) **34**(1), 1–47 (2002)
32. Smucker, M.D., Kazai, G., Lease, M.: Overview of the TREC 2012 crowdsourcing track. In: Voorhees, E.M., Buckland, L.P. (eds.) The Twenty-First Text REtrieval Conference Proceedings (TREC 2012). National Institute of Standards and Technology (NIST), Special Publication 500–298, Washington, USA (2013)
33. Soboroff, I., Nicholas, C., Cahan, P.: Ranking retrieval systems without relevance judgments. In: Kraft, D.H., Croft, W.B., Harper, D.J., Zobel, J. (eds.) Proceeding of 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001), pp. 66–73. ACM Press, New York (2001)
34. Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Inf. Process. Manage. **45**(4), 427–437 (2009)
35. Voorhees, E.M.: Variations in relevance judgments and the measurement of retrieval effectiveness. Inf. Process. Manage. **36**(5), 697–716 (2000)