Benchmarks for Digital Preservation tools

Kresimir Duretec [1], Artur Kulmukhametov[1], Andreas Rauber[1] and Christoph Becker[1,2]

[1]Vienna University of Technology, Austria

[2]University of Toronto, Canada

ABSTRACT

Creation and improvement of tools for digital preservation is a difficult task without an established way to assess any progress in their quality. This happens due to low presence of solid evidence and a lack of accessible approaches to create such evidence. Software benchmarking, as an empirical method, is used in various fields to provide objective evidence about the quality of software tools. However, the digital preservation field is still missing a proper adoption of that method. This paper establishes a theory of benchmarking of tools in digital preservation as a solid method for gathering and sharing the evidence needed to achieve widespread improvements in tool quality. To this end, we discuss and synthesize literature and experience on the theory and practice of benchmarking as a method and define a conceptual framework for benchmarks in digital preservation. Four benchmarks that address different digital preservation scenarios are presented. We compare existing reports on tool evaluation and how they address the main components of benchmarking, and we discuss the question of whether the field possesses the right combination of social factors that make benchmarking a promising method at this point in time. The conclusions point to significant opportunities for collaborative benchmarks and systematic evidence sharing, but also several major challenges ahead.

General Terms

benchmark, digital preservation, software quality

Keywords

benchmark, digital preservation, software quality

1. INTRODUCTION

The number of different research results developing various preservation tools such as JHove2[1](characterization), Jpylyzer[2](quality assurance), Fido[3](identification) and others indicate their importance to the preservation community. The high quality of those tools is of major importance to the community. Although the community tends to acknowledge that better tools are still needed[4], proper evidence to support quality claims is still missing. This makes it hard to quantify the extent to which better tools are needed and how good the current ones actually are. Furthermore, the missing evidence puts major constraints on the decision making procedures which are implemented in various memory institutions.

Evidence, and the lack of it, has been a major concern in several fields closely related to digital preservation. Scientists have argued for experimentation, a type of empirical study, as an important method for providing evidence in software engineering and computer science [4][40]. However, different communities have shown different levels of acceptance of experimentation pointing to numerous reasons, such as costs and challenge to control all the variables, as a limiting barrier for rigorous adoption [40][26]. To address the barriers approaches such as testbeds and benchmarks have been proposed[26][3]. A benchmark is defined as "a standard against which measurements or comparisons can be made"[2]. A testbed is defined as "an environment containing the hardware, instrumentation, simulators, software tools, and other support elements needed to conduct a test"[2]. Even though both methods have comparison of software artefacts as their main goal slight difference can be distinguished. While a benchmark defines how the comparison should be done, a testbed is focused on providing a complete infrastructure to support that comparison. Tichy argued that benchmarks are an effective and affordable way to conduct experiments, although their development can require significant resources[40] .

In the digital preservation field, the term benchmark has been used several times but generally not accompanied by a rigorous treatment of the underlying assumptions, theories, requirements, limitations and techniques that are needed to make effective use of this method. This has resulted in several approaches which have not received sustained follow up. Benchmarks are thus still on the margin in the digital preservation field, even though this method has shown

[1]https://bitbucket.org/jhove2/main/wiki/Home

[2]http://jpylyzer.openpreservation.org/

[3]https://github.com/openpreserve/fido

[4]http://openpreservation.org/blog/2012/10/19/practitioners-have-spoken-we-need-better-characterisation/

major benefits in other fields.

The main contribution of this paper is the introduction of systematic and theory-based benchmarks to the digital preservation field. To enable the community to systematically define, use and evaluate benchmarks, a common model is required to define main benchmark components. Since the development of software tools is the focus, such a model should be based on theories from the software engineering field. Quality aspects of interest to our domain need to be backed up by well-defined quality models and metrics to enable objective comparison of the tools being benchmarked. Authenticity, as a key aspect of digital preservation, points to the correctness of tools as a crucial aspect of quality. However, this aspect has received insufficient effort so far[5]. Although the digital preservation community still lacks these benchmarks, several indicators signify the community's readiness.

This paper is organized as follows. In order to establish the basis for defining the common model, Section 2 provides an overview of the theory and practice of benchmarks in the software engineering and information retrieval fields. This is followed by an overview of related initiatives in the digital preservation field. Section 3 provides a common model for benchmarks. It defines the five main components of each benchmark. Section 4 provides four benchmarks which are described in terms of the five main components defined by the common model. Section 5 discusses the impact of proposed benchmarks and points to several preconditions which indicate community readiness for such benchmarks. Finally Section 6 summarizes the main conclusions and points to the future work.

2. THEORY AND PRACTICE OF BENCHMARKING

2.1 Benchmarks in related fields

The software engineering and information retrieval fields can be identified as most relevant fields for building benchmarks for digital preservation tools. One of the concerns of software engineering is to research and provide methods for evaluating software artefacts. Sim et al.[35]define a benchmark as "a test or set of tests used to compare the performance of alternative tools or techniques". Benchmarking has been a method employed by various laboratories and industries to objectively evaluate software solutions. The information retrieval field is mainly concerned with providing models and methods for an efficient information extraction from different sources. Digital preservation relies heavily on the metadata extracted from digital objects. This extraction, often performed by characterization tools, can also be considered to be a type of information retrieval.

Over the years research communities in software engineering and information retrieval have adopted and further developed benchmarks as a rigorous method to provide empirical evidence. This has provided an additional boost to the research and innovation in those fields. The Transaction Processing Council (TPC)[5] has been releasing a series of benchmarks covering various transaction actions. They have released over 750 benchmark publications covering a range of hardware and software platforms but have become most widely known for their database-centric benchmarks. The information retrieval field has several successful initiatives such as TREC[6], CLEF[7], MediaEval[8], and Mirex[9]. The Text Retrieval Conference (TREC), launched in 1992, has been releasing a number of information retrieval tasks organized in tracks to support evaluation of different information retrieval methodologies (in 2014 eight different tracks were organized). Numerous financial and nonfinancial benefits have been reported. It has been estimated that 16 million dollars of investment in TREC has resulted in 81 million dollars of extrapolated benefits[38]. The nonfinancial benefits are even more impressive ranging from providing large test collections and robust evaluation methodologies to enabling a competition which has fostered the whole research area. Many of the solutions have been adopted by the industry.

2.2 Components of a benchmark

In the software engineering field Sim et al.[35] propose a theory which views benchmarks as social and technical artefacts arising as the result of a consensus in a well-established community. Their interest is focused mainly on the technical research community. They have identified three major benchmark components: motivating comparison, task sample and performance measures, leaving open the order in which those components are developed.

- The **motivating comparison** defines the comparison to be done and the benefits that comparison will bring in terms of the future research agenda. For example, Kienle and Sim [19] motivate their benchmark for fact extraction from web sites by enabling the comparison of capabilities of different fact extractors. Heckman and Williams [16] propose a benchmark for tools that detect anomalies in source code. The main motivation is to find tools with the best rate of anomaly detection.
- The **task sample** is a list of tests that the subject, to which a benchmark is applied, is expected to solve. Kienle and Sim[19] use both artificial and real web sources as task samples for their web site extractors. Heckman and Williams[16] divide their task sample into two parts: six real Java subject programs and a list of true and false anomalies in those programs.
- The **performance measures** are qualitative or quantitative measurements taken by a human or a machine to calculate how fit the subject is for the task. For instance, Heckman and Williams[16] provide a list of well-established measures from the area of data mining and software anomaly detection.

In the information retrieval field Dekhtyar et al.[12] provide five main benchmark components: data set, tasks, answer set, measures and data representation formats/supplementary software. While tasks and measures are similar to the task samples and the performance measures proposed by Sim et al.[35], the typical usage scenario of information retrieval methods has identified data sets with accompany-

[5] http://www.tpc.org/information/about/abouttpc.asp

[6] http://trec.nist.gov/

[7] http://www.clef-initiative.eu/

[8] http://www.multimediaeval.org/

[9] http://www.music-ir.org/mirex/wiki/MIREX_HOME

ing answer sets as important benchmark components. The dataset contains information which a certain tool is required to retrieve. The answer set (often referred to as a ground truth) contains the correct answers which a tool is expected to return. It is reported by several authors that establishing a high-quality ground truth is the biggest challenge of such benchmarks[9][8] and the lack of it is a serious limiting factor [12]. Ben Charrada et al.[8] provide a real-world test dataset for which the ground truth is manually created. To reduce the impact of potential biases which could affect ground truth Chen et al.[9] proposed generating ground truth by a group of participants in several stages. Each stage is supposed to resolve conflicts from the previous stage.
The type of the data(tests) used in a benchmark plays an important role. Seng et al.[34] divide their database system benchmarks into two categories: synthetic and empirical. Synthetic benchmark create artificial data and tests, and empirical benchmarks use real world data and tests. They acknowledge that empirical benchmarks, even though ideal, in the case of databases are prevailed by synthetic due to the lower costs of implementing the synthetic ones.

To evaluate the benchmark quality several authors have proposed a list of desired characteristics. Sim et al.[35] propose a list of seven properties of successful benchmarks. Those are accessibility, affordability, clarity, relevance, solvability, portability and scalability. Huppler[17] proposes a list of five characteristics: relevant, repeatable, fair, verifiable and economical. He stresses repeatability as an important criterion allowing interested parties to get the same result even after repeating the whole benchmark. This criterion contributes to the overall trust in the results provided by a benchmark.

The provided components have shown to be beneficial in both fields as they allowed researchers to provide more focused benchmarks. It is clear that providing a common structure makes easier definition and comparison between similar benchmarks.

2.3 Awareness of the digital preservation community

The NDSA National Agenda for Digital Stewardship 2015 [27] highlights the importance of repeatable case studies and experiments, which are eventually to be transformed into "production public test beds" and "conformance tests". The authors highlight that digital preservation is missing systematic metrics and measurements for "even simple failure scenarios", which are dedicated to bit preservation.

To our knowledge, the first mention of the problem of lacking benchmarking in digital preservation is dated to 2000, when Greenstein[15] identifies benchmarking as an upcoming challenge for digital libraries. One of the early initiatives to create testbeds was carried out within the project Testbed Digitale Bewaring (Dutch Digital Preservation Testbed)[30] in 2002. The aim was to create testbeds for controlled experiments on preservation approaches (migration, emulation, XML) which were planned to be used by the Dutch government. As an example, the authors consider migration of MS Word documents within the testbed. They were interested to study documents features that change during the migration process. During the same time period, the development of testbeds was a key component of the US Digital Library Initiative (DLI) which led to the development of the D-Lib Test Suite[24].

The next milestone was the DELOS Digital Preservation Testbed, created in the DELOS project[37] in 2006. This testbed was based on the Dutch Digital Preservation Testbed. It contained a workflow of 14 steps, which were introduced to simplify the process of benchmarking, to guide users and to automate collection of evidence and documentation.

In 2007, Neumayer et al.[28] describe a range of issues arising when creating a testbed for digital preservation based on the accumulated experience and knowledge in the DELOS project. The challenges were (1) precise task definition, (2) definition of "sufficient" size of a benchmark, (3) benchmark samples generation, (4) data representation, and (5) ground truth and evaluation criteria specification. The authors attempted to empirically generalize on requirements and criteria, fleshing out a common structure of a benchmark.

Creation of the Planets Testbed[25] was inspired by the work undertaken by Dutch and DELOS testbeds in 2010. One of the critiques of the previous works was reliance on manual processes when characterising objects for a testbed. It is a time-consuming and error-prone activity, which is hardly applicable to large collections. The testbed here did not represent an actual real-world setting, but a software environment to explore with, test, and compare preservation tools and services in an online environment. These were open-ended tests, not necessarily focused on performance measures used for ranking tools. In parallel, the well-known decision support tool Plato for preservation planning process was developed[6]. In Plato, the focus is on systematic evaluation for the purpose of ranking and selection, and a strong emphasis is put on measuring and controlling the environment variables that influence results[7]. This makes the experiments rigorous, but the focus is situated on the particular decision making environment of one organization, and the requirements are tailored to these specific needs.

In 2011, the SCAPE project continued the work done on Plato in Planets, but adopted a different approach on the creation of the test environment. The project used its partners as sources for testbeds which were addressed by scenarios and constitute triplets of the following concepts: a dataset, a preservation issue and a possible solution[13]. This allowed them to structure the testbeds and think of potential scenarios and use cases, with limits on generalizability. Although the process of generation of datasets was automated, there is no confidence that the ground truth was valid and correct.

This issue is being addressed in the BenchmarkDP project. It is developing an approach to create benchmark datasets for objective validation of properties, such as functional correctness, of preservation tools [5]. Moreover, this approach allows automated generation of evidence for validity of datasets and corresponding ground truth.

2.4 Observations

As discussed earlier in this section, although there have been initiatives to address some specific cases for benchmarking, a

holistic analysis of this challenge or at least an explicit list of benchmarks required in the digital preservation community does not yet exist. The work performed by the projects in digital preservation is lacking theoretical grounding (such as by Sim et al.[35]), so it is hard to rigorously evaluate requirements and criteria and study limitations of the testbeds.

Despite the existing efforts to create benchmarks and testbeds, there is still a deficiency of tests in digital preservation[33]. Hutchins[18] provided a thorough report on testing characterization tools. He confirms an issue of lacking ground truth datasets and methods, which would make it possible to verify correctness of a standalone tool. Rosenthal[31] also mentions lack of benchmarks in the bit preservation domain. He proposes strategies to improve competition in the market of software tools for bit preservation. The strategies are (1) agreement on common metrics, (2) consensus on modeling techniques for the metrics, (3) generation of better data and metadata, and (4) decreasing human factors as a reason for data loss. These strategies are applicable to the case of digital preservation as well: there is neither any agreement on metrics, nor ways to model these metrics, nor common approaches to create data for benchmarks.

These limitations prevent rigorous testing of the produced software tools. The community is aware of the shortcomings and define them as challenges in the research agenda. Practitioners are becoming aware of potential issues of selecting proper, trustworthy and correct components during decision making.

The benchmarking theory and practices from the other domains explained in this section are the foundations of the proposed approach to create benchmarks. The theory by Sim et al.[35] on benchmarks is a crucial pivot around which the body of benchmarks is to be built. It provides all necessary concepts and models which link the concepts and properties of successful benchmarks.

3. BENCHMARKS IN DIGITAL PRESERVATION

This section proposes a common benchmark model for digital preservation. The digital preservation tool benchmark defines a standardized way to objectively compare various software tools relevant to the digital preservation community. The common benchmark model defines five major components that each benchmark should define. As the focus of this paper is software tools, the model is not meant to be applicable to other areas of digital preservation where benchmarks might be used (e.g. organization benchmarks).

3.1 A common model for benchmarks in digital preservation

The theoretical work proposed by Sim et al[35] forms the basis for the common benchmark model. Based on the three proposed components (motivating comparison, task sample, performance measures) and the importance of data to the digital preservation community, five main benchmark components are identified: (1) motivating comparison, (2) function, (3) dataset, (4) ground truth (optional), and (5) performance measures.

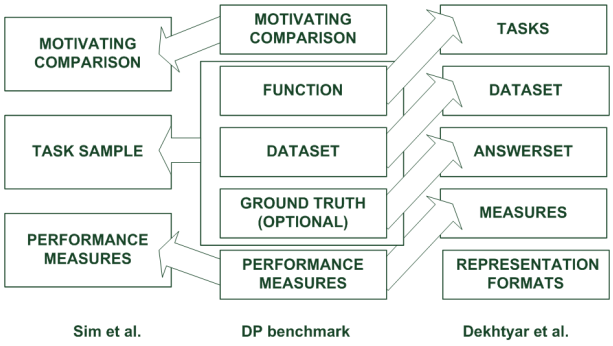

Figure 1: The common benchmark model mapped to the models from the SE and IR community

The *motivating comparison*, as defined by Sim et al. [35], will provide details on what a benchmark is supposed to compare. This can cover a variety of scenarios such as comparing tools in calculating significant properties values from electronic records, comparing different PDF validators or comparing different web harvesters in harvesting web pages. Each benchmark should motivate a comparison which is important to the community and is expected to further the whole research field.
The task sample proposed by Sim et al[35] has been divided into three parts: *function*, *dataset* and *ground truth*.
Function defines a specific task. It can range from migrating an object from one format to another to calculating values of a specific set of properties from a digital object.
The dataset defines a set of digital objects on which the specified task is to be executed. The dataset can be a set of images or documents, but also a set of software components (e.g. a set of video games which might be used in different emulation environments). To enable credible evaluation, in some cases the dataset might be accompanied by an appropriate ground truth.
The ground truth contains correct answers that a certain tool is expected to produce. For some motivating comparisons and task samples this element will not be required.
Performance measures demonstrate the fitness of the benchmarked tool for a certain task. As proposed by Sim et al[35], those measures can be quantitative or qualitative and can be calculated by a human or a machine. Performance measures are benchmark-specific which requires for each benchmark to properly document them together with the criteria for selecting them.

The common benchmark model can be unambiguously mapped to the models proposed in the software engineering[35] and information retrieval[12] fields (Figure 1).

3.2 What to compare and how to measure it? Quality modeling and performance measures

The main goal of the motivating comparison is to provide details on what a benchmark is supposed to compare. This can include various aspects such as the speed of a tool, usability or correctness of output. These quality aspects should be backed up by a quality model to avoid any misinterpretations and improve the clarity of a benchmark.

Table 1: A simple scenario mapped to the common model

Element	Question	Example
Motivating comparison	What to compare?	Correctness of characterization tools when extracting text from files.
Function	Which function?	Extraction of text from files.
Dataset	Which dataset?	MS Word files.
Ground truth	What is the ground truth?	Text inserted into each MS Word file.
Performance measures	What is calculated?	Percentage of files where text was correctly extracted.

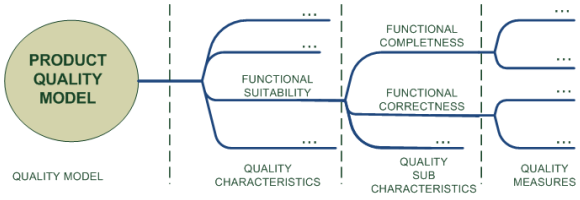

Figure 2: Hierarchical quality model

The ISO SQUARE Product Quality Model [1] organizes quality aspects such as speed, usability, and correctness into eight quality characteristics which are further divided into subcharacteristics. The software quality characteristics and subcharacteristics are indicated by one or more software quality measures[1].

Figure 2 shows the hierarchical decomposition of the Product Quality Model with the most relevant characteristics pointed out. Authenticity, the key concern of digital preservation, is considered when deciding on the relevance of characteristics. The concern is defined as "a degree to which a person or a system regards an object as what it is purported to be"[39]. Various tools capable of manipulating digital objects (e.g. migration) or measuring the values of object properties have the biggest impact on the authenticity of the digital object. Arguably the correctness of such tools is the most important quality aspect. The characteristic Functional Suitability and its two subcharacteristics Functional Completeness and Functional Correctness are identified as the most important characteristics related to authenticity. Those cover the degree to which a certain tool covers all the needed tasks and produces correct results[1].

The mentioned quality model provides a link between the motivating comparison and performance measures. The link is helpful to validate a selection of measures that are used to address a tool's quality. This will contribute to the clarity of benchmark specifications. An example of such linking is shown in Table 1 where the characterization tool's correctness is indicated as a percentage of files where the characterization task was successful.

As acknowledged by Sim et al. [35], creating performance measures (software quality measures) is particularly difficult. The digital preservation field has a systematic list of relevant quality measures [10] based on an ontology[23]. To expand this the information retrieval field with its numerous quality measures can be considered[36].

4. A SET OF BENCHMARKS

In this section, two benchmarks are introduced in details to demonstrate the applicability of the theory proposed previously. Additionally, there is a description of other benchmarks in Table 3. The proposed benchmarks are composed of the five components defined by the common model in Section 3.1. It is expected that each benchmark satisfies desired qualities defined by Sim et al[35]. However, due to the limited space the main focus of discussion is the relevance of the proposed benchmark to the digital preservation community and affordability. Thus the main goal of each benchmark definition is to provide a clear motivation to the digital preservation community and an understanding of what the benefits would be to the community when the benchmark is created and used. Furthermore, each benchmark will provide a clear and concise overview of the main tool function to be compared, requirements for the dataset, the nature and structure of the ground truth and an overview of applicable performance measures. Finally, each benchmark specification discusses major challenges that are expected when implementing the benchmark.

4.1 Raw photograph migration to DNG

4.1.1 Introduction

Raw photographs are images made by cameras and stored in a raw format. When considering digital preservation of raw photographs[22], migration is the most suitable candidate. This approach helps to avoid a risk of information loss due to discontinuation of support from a manufacturer. There are currently many proprietary raw formats with an undetermined lifetime. A common strategy to overcome this issue is migration to an open-source and standardized format. In this case, the format is DNG (Adobe Digital Negative). There are tools that allow such migrations like Photoshop[11], DNG Converter[12], CaptureOne[13], DigiKam[14] etc. Their application promises operational short-term benefits of homogeneous datasets that are easier to manage, as well as long-term benefits of lower risks of losing access to the assets. However, usually there is no evidence or confirmation based on rigorous testing that the tools work correctly during execution of a migration process. Therefore, the tools are not trustworthy and using them in preservation operations is risky. This benchmark enables the ranking of migration tools against a dataset of raw photographs. This is a practical problem for professionals and institutions, who consider selecting the best tool for raw photographs.

4.1.2 Motivating comparison

The purpose of this benchmark is a comparison of correctness of migration processes done by various software tools on the photograph dataset. It will show how similar a migrated is to the original photograph in terms of an image content,

[10] purl.org/dp/quality
[11] http://www.adobe.com/
[12] http://www.adobe.com/
[13] http://www.phaseone.com/
[14] http://www.digikam.org/

not metadata. Kulmukhametov et al.[22] discussed technical challenges when calculating similarity of photographs and introduced a tool, which will be used in this benchmark. The tool implements an algorithm which calculates Structure Similarity (SSIM) measure, which is claimed to be the closest measure to human perception when considering similarity of two images.

4.1.3 Function

The function to benchmark is migration of photographs from proprietary raw formats to the DNG format. As the files store raw data, tools usually do not provide many adjustable parameters, which would affect the content. Image compression is the only setting provided by the migration tools. As the goal of this benchmark is to test a correctness of migration, compression may significantly reduce the overall quality of the resulting photograph. This feature must be turned off.

4.1.4 Dataset

The dataset consists of photographs stored in raw formats. As the task is to compare correctness of migration tools, the dataset consists of photographs produced by different cameras and manufacturers. Such a dataset allows one to rank tools and determine the most versatile and universal one. Populating the dataset with photographs is achievable by using a content profiling tool C3PO[21], which allows one to extract samples from a bigger collection based on a specified list of criteria: a raw format, a camera model, a manufacturer.

4.1.5 Performance measures

The correctness of the tool is measured by calculating the SSIM value of a migrated photograph. The value is measured from 0 to 1. A higher magnitude of the value means better results. It is possible to compare the values from different tools for one photograph of the dataset. This makes it possible to identify the best tool for migration of this digital object. Another possibility is to calculate statistics based on the results of running the migration process for the whole dataset by one tool. The statistics, such as mean, median and standard deviation, may be helpful to identify the most versatile tool which produces the best results for the dataset.

4.1.6 Discussion

There is a challenge associated with this benchmark. It is about which photographs will constitute the dataset. There is no simple answer as the population of photographs is unknown. One possible solution is to provide samples of photographs created by different cameras from different manufacturers. Focusing on specific situations based on the requirements of the community is an important contribution to solve this challenge.

4.2 Property extraction from documents in electronic records environments

4.2.1 Introduction

Electronic records cover a spectrum of different use cases such as emails, audio or video records or documents. Document-based electronic records furthermore can cover a variety of scenarios such as books, articles or contracts.

In many of those scenarios, document authenticity is of key importance. A migration tool can affect authenticity of a document by falsely migrating or not migrating at all some document elements. The lack of proper evidence around these cases makes it challenging to demonstrate authenticity of a document created by a migration.
To provide evidence for document authenticity, values of various document properties are measured. Pairs of property and value form a characteristic [11]. Stakeholders often point to significant properties of a document as important for its authenticity[29]. Expressing those properties in a measurable form enables assertion of document authenticity.

A number of different characterization tools, such as Apache Tika[15], National Library New Zealand Metadata Extractor[16] or Jhove2[17] claim to be capable of measuring values of various document properties. As they cover the commonly used formats such as MS Word and PDF they are suitable for providing evidence that is important for document authenticity. However, the coverage of needed properties and the correctness of measured values is not fully covered by a rigorous evaluation. This still hampers the validation of document authenticity as it is not possible to establish the confidence in the measured values.

Therefore, a benchmark is proposed to enable a rigorous evaluation of characterization tools when measuring document property values.

There are several major benefits of such a benchmark. The most important benefit is that it would provide the needed evidence around the quality of different characterization tools and enable an objective comparison of them. Furthermore, it is expected that it would foster the future development of those tools which would lead to better characterization tools. This would also be beneficial for establishing proper migration benchmarks which would be able to rigorously evaluate migration tools. As highlighted by Ross, "before we can see migration as a viable aid to preservation, more work is needed in the development of metrics for benchmarking and supporting the evaluation of the risks or losses resulting from particular changes"[32].

4.2.2 Motivating comparison

The purpose of this benchmark is to enable the comparison of characterization tools with respect to the coverage of document properties and correctness of measured values for those properties. Coverage can be mapped to the functional completeness quality characteristic and correctness to the functional correctness. The Functional completeness is included mainly to denote if a certain tool can measure a property value. It is expected that in some cases some properties will not be fully covered which makes it an even more important aspect to systematically evaluate and compare.

4.2.3 Function

The main function is measuring values of document properties. Due to their importance for the authenticity, significant

[15]http://tika.apache.org/
[16]http://meta-extractor.sourceforge.net/
[17]https://bitbucket.org/jhove2/main/wiki/Home

Table 2: Quality characteristics and performance measures

Quality Characteristic	Measure	Calculated as
Functional completeness	Coverage	calculated per property as a percentage of documents where a tool returned a value for a specific property
Functional correctness	Accuracy	calculated per property as a percentage of files where a tool returned correct value for a specific property
	Exact Match Ratio	calculated for a tool as a percentage of files where the tool returned correct values for all properties

properties are in the focus of the benchmark. However as pointed out by Dappert et al.[11] the significance of a property is not absolute and binary but depends on the stakeholders' requirements for a certain document(or a collection of documents). Thus it will be challenging or even impossible to come up with a list of required significant properties. However, it can be argued, due to the similar scenarios various content holders are dealing with, that it is possible to come up with a list of commonly used properties which are identified as significant for documents in electronic records environments. Building on previous studies that classified and modelled significant properties in preservation planning case studies, and by analyzing actual preservation plans created by different stakeholders a list of common significant properties can be made.

4.2.4 Dataset and ground truth

In order to cover different documents types the dataset should be focused on the combinations of different document elements and their properties. Here, document elements denote simple building blocks which are used to compose a document (pages, footers, text, images) and their properties such as font color, table size, and image position. This affects the size of the needed dataset. The bigger the number of elements and their properties covered, the bigger the dataset. Even in a very simplistic scenario with five elements where each element has three properties with only one possible value the dataset would need to contain 125 documents to cover all the combinations. The real world is much more complex with more elements, properties and their possible values. This combinatorial explosion makes automatic dataset generation a better method, than the manual annotation, for establishing a proper dataset.

To enhance automation the ground truth needs to be expressed in a machine-readable form. It should specify the correct property-value pairs.

4.2.5 Performance measures

This benchmark addresses two quality characteristics (the functional completeness and the functional correctness). Each characteristic is indicated by one or more measures.

The functional completeness is covered by one measure. This measure should point out how well a single tool covers defined properties. Therefore for each property a percentage is calculated to show the number of documents where a tool returned a value for specific property.

When dealing with functional correctness, there are two aspects that are important to consider. There is the need for a measure that will show how good a tool is on the whole set of properties and on a specific property. For example, it can be important to know that a specific tool which does not have good overall performance has remarkable performance on one of the properties.

4.2.6 Discussion

The proposed benchmark would bring several benefits to the digital preservation community. It would enable an objective comparison of characterization tools in terms of their coverage and correctness when measuring significant properties from documents in electronic records environments. This would provide objective evidence and drive the future development of tools.

The biggest challenge of this benchmark is the dataset generation. Its combinatorial growth, dependent on the number of elements and properties, makes manual annotation insufficient as a method for dataset generation. Automatic dataset generation should provide efficient methods to model different documents in terms of their possible elements and how to control the combinations of those elements and their properties. The model-driven engineering framework[5] provides a possible solution for this problem. The feasibility of the approach has been demonstrated on a similar scenario. However, future work will be required to enhance the whole method to be more robust and cover a larger number of elements.

Once created, it is expected that the effort required for running the benchmark will not be significant. The dataset, even though expected to have a significant number of objects, is still expected to be in the range which standard commodity hardware can handle. Using an artificial dataset raises some issues around the relevance of the benchmark. The biggest challenge that the generation method will need to address is the representativeness of the generated dataset of real-world datasets.

4.3 PDF validation and Web harvesting benchmark

Due to limited space, two additional benchmarks are presented in Table 3. The two benchmarks cover the scenarios of PDF validation and web harvesting.

The PDF file format family has been proliferated over the years as the defacto standard for storing and exchanging various kinds of documents(articles, books, ...). The quality of available validators, used to check the validity of a PDF file, is diverse and hard to objectively compare. Initiatives to build even more validators[18] show that the community is still not satisfied with existing offerings. This points to

[18]http://openpreservation.org/news/verapdfa-consortium-awarded-phase-1-of-preforma-call-for-tender-for-pdfa-validation/

Table 3: PDF validation and Web harvesting benchmarks

Name	PDF validation	Web harvesting
Motivating comparison	Compare validation functional correctness of different PDF file format validators. Furthermore compare the functional correctness of reported violations	Compare functional correctness and completeness of a web harvester
Function	Validate a PDF file	Harvest a web site
Dataset	PDF files covering valid and invalid examples. Invalid examples cover various combinations of violations	A set of webpages. Web-pages are accessed by providing a GET request to a web-server. The settings of the server are set in the benchmark.
Ground truth	Information pointing to the true validity of a PDF file. In the case of an invalid file provides the true violations expected to be reported from a tool	A list of properties for each web-page in the data set: size of the web-page, HTTP GET request, html markup, presence of resources and executable scripts
Performance measures	Accuracy of a validation output; Accuracy of reported violations	Correctness and completeness of the web harvesting tools are measured by calculating precision for the properties

the need for a proper benchmark to enable a proper tool evaluation and comparison. The benchmark would provide an objective evaluation of PDF validation tools.

Web harvesting is an important function in the web archiving community. However due to the complexities of current web pages in terms of links and various technologies being used (e.g. JavaScript and Flash) it is hard to understand the completeness and correctness of the harvesting task.
The proposed benchmark should therefore enable rigorous testing of web harvesting tools by focusing on aspects such as the use of JavaScript, Flash or complex linking structures (spider traps).

5. DISCUSSION

5.1 Preconditions and success factors

Benchmarking as a rigorous method is not a simple, easily completed task. Does our community meet the required preconditions for benchmarking? It is worth revising the requirements and success factors highlighted by Sim [35].
Benchmarks should be collaborative, open, and public. The community has a long track record of sharing various forms of knowledge; however, this has not been replicated when it comes to sharing data. Despite efforts such as LDS3[19], the OPD data endpoint[20], and isolated data sets such as from the UK Web Archive[21], data sharing is not common for a number of reasons. We hope to address some of this by generating data that can be shared freely.

The community must be ready to incur the costs of benchmarking. Continued evolution of the benchmark will be necessary. It will require a selective approach with a focus on those motivating comparisons that are truly encapsulating the paradigms of the field to catalyze substantial interest of the community.

Benchmarks encapsulate paradigms. Benchmarks must be developed by consensus. Are our paradigms understood well enough?

Design decisions need to be supported by lab work. Benchmarking needs to use established results where possible. We base the existing work and proposals in this paper on extensive lab work and case studies in preservation planning and beyond.

Choosing the task sample may be controversial. Consensus is needed in the community, and efforts as part of BenchmarkDP are focused on outreach and community engagement.

The community must have an opportunity to participate, provide feedback, and endorse benchmarks. Efforts should be led by a small number of champions. IPRES as the leading conference in the field is the ideal place for engagement and participation. The authors encourage interested community members to get involved.

5.2 Datasets and ground truth: the key challenge

There is a general lack of test datasets with acompanying ground truth for preservation tools. The widely known and used dataset is the Govodcs dataset[22][14]. However, the only available ground truth is related to identification data. Since that data has been produced by a forensics tool, provided by Forensics Innovations[23] the validity of the ground truth is hard to confirm. Furthermore, the whole dataset is applicable to a limited range of identification scenarios.
Two main approaches in creating test datasets are identified: 1) *subsampling real world datasets and manually annotating them*, and 2) *automatically generating datasets with an accompanying ground truth.*

Manual datasets annotation brings one obvious advantage. Using a real-world test sample makes the benchmark relevant to the real-world scenarios and as such the benchmark results are more trustworthy. However, producing such datasets will be an effort-intensive job and datasets will need to be reduced to a smaller number of objects to make manual annotation plausible. In order to remove any kind of unwanted biases, automatic methodologies for analysing and subsampling real datasets are required. The content

[19] http://beta.lds3.org/
[20] http://wiki.opf-labs.org/display/PT/The+OPF+Data+Endpoint
[21] http://data.webarchive.org.uk/opendata/ukwa.ds.2/
[22] http://digitalcorpora.org/corpora/files
[23] http://www.forensicinnovations.com/

profiling tool C3PO[24] provides a scalable architecture for automatic content profiling of digital objects. It thus, provides the basis on top which sampling algorithms can be built [21][10]. For some functions this kind of sampling from a real world dataset will be sufficient and was the common approach until now in the community.

In some cases e.g. where detailed annotations about technical details are required or the number of features or their combinations require significant size of a sample set and manual annotation might still be too expensive or even impossible. In those situations automatic datasets generation is a possible approach. While in other fields this approach is already researched, the digital preservation field has only started to explore its possibilities and the approach is considered to be highly novel[5][20]. Becker and Duretec [5] proposed a framework based on Model Driven Engineering principles for automatic test dataset generation. This framework has been the basis for several prototypes that serve as a proof of concept. However this is a novel approach in the digital preservation and as such will require significant research effort.

6. CONCLUSION AND FUTURE WORK

Much of the research effort in digital preservation is invested in developing software tools for managing, processing and disseminating digital information. The community has increasingly recognized the need for systematic testing and evidence sharing on different characteristics of quality of those tools. In this article, we introduced insights from theory and practice of benchmarking of Software Engineering and Information Retrieval communities and discussed how the introduction of systematic benchmarking provided a boost for research and innovation in these communities. Based on a simple framework for specifying and analyzing benchmarks, we outlined a set of initial specifications for benchmarks. While this initial set is by no means complete, it provides a key stepping stone towards collaborative campaigns for benchmarking. The defined four benchmarks will be a starting point for community involvement in establishing benchmarking in digital preservation as an important method for strengthening the evidence base.

An essential characteristic of a successful benchmark is that it will lead to better tools, to the point that a majority of tools complete standard benchmarks with near-perfect scores. This means that it is possible to start with quick wins for comparison tasks that are comparably simple, but relevant for comparison, roadmap generation and prioritization of future development of tools, in order to establish the mechanisms of benchmarking as a method; and then proceed to advanced, more challenging benchmarks as experience accumulates.

But more importantly, it means that each successful benchmark will eventually be superseded by an evolved specification. It will require joint community interest and efforts to make such efforts feasible and worthwhile; and hence, a focus is needed on those quintessential tasks for which a systematic, rigorous comparison of candidate components on a widely agreed performance measure is possible, necessary, and relevant. It is up to the members of the community to ensure that their needs are part of this consensus.

[24]http://ifs.tuwien.ac.at/imp/c3po

Acknowledgements

Part of this work was supported by the Vienna Science and Technology Fund (WWTF) through the project *BenchmarkDP* (ICT12-046).

7. REFERENCES

[1] *ISO/IEC 25010 Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models*. 2010.

[2] Systems and software engineering – Vocabulary. *ISO/IEC/IEEE 24765:2010(E)*, pages 1–418, Dec. 2010.

[3] E. Barreiros, A. Almeida, J. Saraiva, and S. Soares. A Systematic Mapping Study on Software Engineering Testbeds. In *2011 International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pages 107–116, Sept. 2011.

[4] V. Basili. The role of experimentation in software engineering: past, current, and future. In *, Proceedings of the 18th International Conference on Software Engineering, 1996*, pages 442–449, Mar. 1996.

[5] C. Becker and K. Duretec. Free Benchmark Corpora for Preservation Experiments: Using Model-driven Engineering to Generate Data Sets. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '13, pages 349–358, New York, NY, USA, 2013. ACM.

[6] C. Becker, H. Kulovits, A. Rauber, and H. Hofman. Plato: A service oriented decision support system for preservation planning. In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, pages 367–370. ACM, 2008.

[7] C. Becker and A. Rauber. Improving component selection and monitoring with controlled experimentation and automated measurements. *Information and Software Technology*, 52(6):641–655, 2010.

[8] E. Ben Charrada, D. Caspar, C. Jeanneret, and M. Glinz. Towards a Benchmark for Traceability. In *Proceedings of the 12th International Workshop on Principles of Software Evolution and the 7th Annual ERCIM Workshop on Software Evolution*, IWPSE-EVOL '11, pages 21–30, New York, NY, USA, 2011. ACM.

[9] X. Chen, J. Hosking, J. Grundy, and R. Amor. Development of Robust Traceability Benchmarks. In *Software Engineering Conference (ASWEC), 2013 22nd Australian*, pages 145–154, June 2013.

[10] Christoph Becker, Luis Faria, and Kresimir Duretec. Scalable decision support for digital preservation. *OCLC Systems & Services: International digital library perspectives*, 30(4):249–284, Nov. 2014.

[11] A. Dappert and A. Farquhar. Significance is in the Eye of the Stakeholder. In *Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL'09, pages 297–308, Berlin, Heidelberg, 2009. Springer-Verlag.

[12] A. Dekhtyar and J. Hayes. Good Benchmarks are Hard To Find: Toward the Benchmark for Information Retrieval Applications in Software Engineering. *Information Retrieval in Software Engineering, International Conference on Software Maintenance (ICSM): Philadelphia, PA.*, Sept. 2006.

[13] M. Ferreira, H. Silva, R. Castro, P. Moldrup-Dalum, Z. Pehlivan, C. Wilson, and S. Schlarb. D10.2 gap analysis on action services tools and scape platform and testbeds requirements, 2013.

[14] S. Garfinkel, P. Farrell, V. Roussev, and G. Dinolt. Bringing Science to Digital Forensics with Standardized Forensic Corpora. *Digital Investigation*, 6:S2–S11, Sept. 2009.

[15] D. Greenstein. Digital libraries and their challenges. *Library trends*, 49(2):290–303, 2000.

[16] S. Heckman and L. Williams. On Establishing a Benchmark for Evaluating Static Analysis Alert Prioritization and Classification Techniques. In *Proceedings of the Second ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, ESEM '08, pages 41–50, New York, NY, USA, 2008. ACM.

[17] K. Huppler. The Art of Building a Good Benchmark. In R. Nambiar and M. Poess, editors, *Performance Evaluation and Benchmarking*, number 5895 in Lecture Notes in Computer Science, pages 18–30. Springer Berlin Heidelberg, 2009.

[18] M. Hutchins. Testing software tools of potential interest for digital preservation activities at the national library of australia. *National Library of Australia Staff Papers*, 2012.

[19] H. Kienle and S. Sim. Towards a benchmark for Web site extractors: a call for community participation. In *Seventh European Conference on Software Maintenance and Reengineering, 2003. Proceedings*, pages 82–87, Mar. 2003.

[20] Y. Kim and S. Ross. Searching for Ground Truth: A Stepping Stone in Automating Genre Classification. In C. Thanos, F. Borri, and L. Candela, editors, *Digital Libraries: Research and Development*, number 4877 in Lecture Notes in Computer Science, pages 248–261. Springer Berlin Heidelberg, 2007.

[21] A. Kulmukhametov and C. Becker. Content Profiling for Preservation: Improving Scale, Depth and Quality. In K. Tuamsuk, A. Jatowt, and E. Rasmussen, editors, *The Emergence of Digital Libraries – Research and Practices*, number 8839 in Lecture Notes in Computer Science, pages 1–11. Springer International Publishing, Nov. 2014.

[22] A. Kulmukhametov, M. Plangg, and C. Becker. Automated quality assurance for migration of born-digital images. In *Archiving Conference*, volume 2014, pages 73–78. Society for Imaging Science and Technology, 2014.

[23] H. Kulovits, M. Kraxner, M. Plangg, C. Becker, and S. Bechhofer. Open preservation data: Controlled vocabularies and ontologies for preservation ecosystems. *Proc. IPRES*, pages 63–72, 2013.

[24] R. L. Larsen. The dlib test suite and metrics working group: Harvesting the experience from the digital library initiative. *D-Lib Working Group on Digital Library Metrics Website*, 2002.

[25] A. Lindley, A. N. Jackson, and B. Aitken. A collaborative research environment for digital preservation-the planets testbed. In *Enabling Technologies: Infrastructures for Collaborative Enterprises (WETICE), 2010 19th IEEE International Workshop on*, pages 197–202. IEEE, 2010.

[26] M. Lindvall, I. Rus, F. Shull, M. Zelkowitz, P. Donzelli, A. Memon, V. Basili, P. Costa, R. Tvedt, L. Hochstein, S. Asgari, C. Ackermann, and D. Pech. An evolutionary testbed for software technology evaluation. *Innovations in Systems and Software Engineering*, 1(1):3–11, Mar. 2005.

[27] National Digital Stewardship Alliance. 2015 National Agenda for Digital Stewardship, 2015.

[28] R. Neumayer, H. Kulovits, M. Thaller, E. Nicchiarelli, M. Day, H. Hofmann, and S. Ross. On the need for benchmark corpora in digital preservation. In *Proceedings of the 2nd DELOS Conference on Digital Libraries*, 2007.

[29] Parliamentary Archives. *A Digital Preservation Policy for Parliament*. London, Parliamentary Archives, 2009.

[30] M. Potter. Researching long term digital preservation approaches in the dutch digital preservation testbed (testbed digitale bewaring). *RLG DigiNews*, 6(3), 2002.

[31] D. S. Rosenthal. Bit preservation: A solved problem? *International Journal of Digital Curation*, 5(1):134–148, 2010.

[32] S. Ross. Changing Trains at Wigan: Digital Preservation and the Future of Scholarship, Jan. 2000.

[33] R. Ruusalepp and M. Dobreva. Digital preservation services: State of the art analysis. 2012.

[34] J.-L. Seng, S. B. Yao, and A. R. Hevner. Requirements-driven database systems benchmark method. *Decision Support Systems*, 38(4):629–648, Jan. 2005.

[35] S. E. Sim, S. Easterbrook, and R. C. Holt. Using Benchmarking to Advance Research: A Challenge to Software Engineering. In *Proceedings of the 25th International Conference on Software Engineering*, ICSE '03, pages 74–83, Washington, DC, USA, 2003. IEEE Computer Society.

[36] M. Sokolova and G. Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, July 2009.

[37] S. Strodl, A. Rauber, C. Rauch, H. Hofman, F. Debole, and G. Amato. *The DELOS testbed for choosing a digital preservation strategy*. Springer, 2006.

[38] G. Tassey, B. R. Rowe, D. W. Wood, A. N. Link, and D. A. Simoni. Economic Impact Assessment of NIST's Text REtrieval Conference (TREC) Program. *National Institute of Standards and Technology, Gaithersburg, Maryland*, 2010.

[39] The Consultative Committee for Space Data Systems. *Reference Model for an Open Archival Information System (OAIS)*. June 2012.

[40] W. Tichy. Should computer scientists experiment more? *Computer*, 31(5):32–40, May 1998.