Free Benchmark Corpora for Preservation Experiments: Using Model-Driven Engineering to Generate Data Sets

Christoph Becker
Vienna University of Technology
Vienna, Austria
www.ifs.tuwien.ac.at/~becker

Kresimir Duretec
Vienna University of Technology
Vienna, Austria
www.ifs.tuwien.ac.at/~duretec

ABSTRACT

Digital preservation is an active area of research, and recent years have brought forward an increasing number of characterisation tools for the object-level analysis of digital content. However, there is a profound lack of objective, standardised and comparable metrics and benchmark collections to enable experimentation and validation of these tools. While fields such as Information Retrieval have for decades been able to rely on benchmark collections annotated with ground truth to enable systematic improvement of algorithms and systems along objective metrics, the digital preservation field is yet unable to provide the necessary ground truth for such benchmarks. Objective indicators, however, are the key enabler for quantitative experimentation and innovation.

This paper presents a systematic model-driven benchmark generation framework that aims to provide realistic approximations of real-world digital information collections with fully known ground truth that enables systematic quantitative experimentation and measurement and improvement against objective indicators. We describe the key motivation and idea behind the framework, outline the technological building blocks, and discuss results of the generation of page-based and hierarchical documents from a ground truth model. Based on a discussion of the benefits and challenges of the approach, we outline future work.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.7 Digital Libraries

Keywords

Repositories, Digital Preservation, Characterisation, Benchmark, Data Set, Ground Truth, Corpora, Model Driven Engineering

JCDL'13, July 22-26, 2013, Indianopolis, Indiana, USA.

General Terms

Algorithms, Experimentation, Measurement, Performance

1. INTRODUCTION

The field of Digital Preservation is concerned with keeping digital information authentic, understandable, and usable, through time and across changing socio-technological environments to achieve digital longevity [19]. Essentially, the fundamental problem addressed is a misalignment of technology: To render information usable to any human, an algorithm needs to produce an interpretation that can be perceived by the human. Digital preservation (DP) is in this sense often seen as a case of interoperability through time. Taking a Shannon communication channel [22] as a metaphor, the core problem is that transmission is asynchronous and may last an indefinite time. At the time of receiving the message, the recipient may not possess an appropriate decoder, the sender may not exist anymore, and the recipient may not be the original addressee. The communication channel thus will often need to convert the message so that the original message intention is preserved. Theoretically, any type converter function carrying out such a transcoding should respect the type of the original message [31]. However, the complexity and change rate of common environments and information representations today have the effect that this is rarely the case.

From its origin in the areas of cultural heritage and eScience, DP has emerged as a key challenge for information systems in almost any domain from eCommerce and eGovernment to manufacturing, finance, health, and private lifestyle. However, the field of DP is approaching a phase in which several key areas of research are hitting a glass ceiling. This applies in particular to the time-intensive area of object- and collection level analysis of existing content, which is an active area of research and development [6, 7, 10, 13, 25, 26, 27].

Authenticity as key requirement for digital preservation requires the converter to prove that the type conversion was respectful. Alternatively, the communication environment needs to provide an appropriate decoder at the time of rendering the message to the receiver. Similar verification is necessary to assure the authentic message is communicated successfully. Unfortunately, current verification methods are often insufficient for this purpose, independent on whether the primary action taken is object migration or environment emulation [2, 13, 5].

There is a profound lack of objective, standardised and comparable metrics and benchmark collections for experimentation. While fields such as Information Retrieval have

for decades been able to rely on benchmark collections annotated with ground truth to enable systematic improvement of algorithms and systems along objective metrics, DP is yet unable to provide the necessary ground truth for such benchmarks. These alone, however, are the key enabler for systematic quantitative experimentation, measurement and improvement against objective indicators.

The current approach to creating benchmarking collections relies on an ex-post analysis of existing content collections with existing tools. However, they produce interpretations with unknown reliability, since prior to exposure to such a computational interpretation, the properties of any digital object are entirely unknown.

This causes a fundamental problem: There is no ground truth that can be used safely to evaluate approaches and system parameters on a large scale [2]. This 'black box phenomenom' differentiates the design problem from scenarios with known ground truth, where key experiment parameters can be explored systematically in large-scale experiments.

In this article, we present a novel approach to the problem of benchmark data sets in preservation. Our hypothesis is that we can turn around the approach towards the construction of such a benchmark from previous ex-post analysis of real-world data to a theory-based, fundamentally model-driven and statistically solid generative approach. To this end, we outline a conceptual framework that combines in-depth statistical profiling of real-world collections with a generative approach based on model-driven engineering. We demonstrate the feasibility of the approach on the example domain of page-based documents. We present results that demonstrate the feasibility of successfully generating of data with fully-known ground truth, and discuss future steps of research and development.

The article is structured as follows. The next section will outline the state of art in characterisation for digital preservation and summarise the challenges that the field community is currently facing. Section 3 motivates and outlines the framework of our approach. Section 4 provides a discussion of the technology stack in use and presents the results that can currently be obtained from the developed dataset generation framework. Section 5 summarises and evaluates the key contributions and outlines next steps ahead.

2. CHARACTERISATION AND BENCHMARKING IN DP

Given how central the notion of data format has been to much of DP research and development, it is not surprising that characterisation has received much attention. A distinction is generally made between *identification* of the format of an object, *validation* of the conformance of the object to the format specification, and *feature extraction* of the object's properties, which is sometimes just called *characterisation*.

The National Archives' DROID tool and siblings such as fido[1], but also the well-known Unix command *file* are typical examples of tools for identification, while tools such as the eXtensible Characterisation Languages XCL [25] and the JSTOR/Harvard Object Validation Environment JHove[2] are probably the most widely cited examples for feature extraction. The File Information Tool Set (FITS)[3] wraps several other tools and combines their output. Following up on the widely used JHove, the JHove2 framework promises to be much improved [1]. However, the improvements are described as "performance improvements and significant new features, most notably, a flexible rules-based assessment capability" [1], and it is unclear whether objective indicators for functionality and functional correctness are used.

While it is clear to the community that in practice, better tools are needed[4], there is much less clarity as to how the quality of these tools is to be tested systematically.

Recent reports from Australia and Europe describe attempts to systematic experimentation. The SCAPE project[5] evaluated the performance and stability aspects of identification tools, touching briefly on aspects such as accuracy and format coverage, using a private data set and manual testing of tools [27]. Aiming at larger scales, the project then ran a number of identification tools against the publicly available Govdocs1 corpus[6]. The 'ground truth' used to evaluate identification accuracy was generated by a forensics tool provided by Forensic Innovations[7] [26]. However, the veridicality of this ground truth is unclear, and the experiment was restricted to identification only.

Others have explored large-scale analysis and aggregation of tool results, as well as visualiation. Tarrant [24] presented a platform for experiment data publication including some visualisation; Jackson discusses longitudinal analaysis of format identification results from the UK web [12], and Petrov demonstrated large-scale aggregation of in-depth characterisation results [17]. However, no objective, trustworthy validation of functional correctness of any of the underlying characterisation tools has been published. Hutchins [10] did include feature extraction in their testing process, but did not attempt to systematically verify the correctness of extracted features. Where indications are found that the results may be incorrect, the conclusion is to not recommend usage of the tool in question.

Arguably, the *functional correctness* of such tools is their most important, crucial quality [3]. It refers to the "degree to which a product or system provides the correct results with the needed degree of precision" [11]. Achieving functional correctness within certain time and resource constraint is a key development challenge, and only by providing solid evidence can we assure full trust in the accuracy of tools. However, little solid evidence exists that describes the functional correctness of any of these tools in objective indicators.

In order to make this critical quality attribute measurable, we need to define meaningful metrics that can be compared objectively, i.e. external measures with unambigously specified semantics. Consider a set of objects $O = \{o_1, o_2, ...o_n\}$ and a set of properties $P = \{p_1, p_2, ...p_m\}$. This might simply be a set of images and the property set {*imageWidth, imageHeight, RGBpixelArray, creatingApplication*}, to use an oversimplified example. The task of feature extraction involves computing the values of each property for each object, which may involve complex transformations. We can

[1] `https://github.com/openplanets/fido`
[2] `http://jhove.sourceforge.net/`
[3] `http://code.google.com/p/fits/`
[4] `http://www.openplanetsfoundation.org/blogs/2012-10-19-practitioners-have-spoken-we-need-better-characterisation`
[5] `http://http://scape-project.eu/`
[6] `http://digitalcorpora.org/corpora/files`
[7] `http://www.forensicinnovations.com/`

express the metrics for measuring correctness directly as the well-known metrics of precision and recall, where

- *Precision* is the fraction of measures obtained that are relevant and within a certain error margin, and
- *Recall* is the fraction of requested measures that are obtained correctly, i.e. within a certain error margin.

Note that we are not concerned with the relevance of each property, which is a contextual decision to be made by the curator, depending on the goals and intents of an organisation [30, 3]. Hence, this specific retrieval problem is independent of the much disputed relevance question asking in which context these properties are deemed significant (cf. [4],[30], and others) and of the applicability of precision and recall to other real-world Information Retrieval problems.

The need for testbed corpora has been a subject of discussion for years [15]. It is clear that annotated benchmark data are needed to support the objective comparison of new approaches and quantify the improvements over existing techniques. However, publicly available data sets are devoid of useful annotations to validate feature extraction, and only limited information is available to provide solid verification for identification. A combination of barriers still prevent the community from actually having these building blocks:

- There are legal constraints on sharing existing data.
- It is technically challenging to develop robust benchmark data.
- There are economic resource constraints on data collection, annotation, sharing, and developing systematic and coherent approaches.
- There is no central reference point or body to coordinate such benchmarking

The technical challenge in this case lies in the well-known, but often forgotten duality of data and computation: In contrast to analog artifacts, any digital artifact is per se a black box, inaccessible to any user. Instead, a computing environment will parse the file, convert it to an internally transformed model and initiate a performance that enables a consumer to conceptualise the *object* (or not, if the performance is inadequate for this goal) [9]. The fundamental issue is one of representation, model equivalency, and interaction: How can information properties be assured across changing environments? Projects such as SCAPE are developing Quality Assurance tools to compare renderings in different environments to produce quantitative measures of equivalence. However, these developments do not address the fundamental obstacle that opposes any such endeavour: The lack of a ground truth that can be used for objective, quantitative evaluation. However, such benchmarking is the irreplaceable milestone that alone enables quantitative improvement and baseline comparison.

Generating ground truth semi-manually is very effort intensive even for tasks such as document image understanding, where the rendering is assumed to be perfect [23]. In areas such as document analysis and recognition, small-scale datasets exist that are created and annotated semi-manually [29, 28]. For page segmentation, a system has been presented that generates images and the corresponding ground truth automatically [8].

To be useful for validating characterisation processes, however, we would need an annotation that maps the conceptual content of objects to their symbol structures used for encoding the digital artifact. This is computationally complex and almost impossible for human experts even for simple file formats. Consequently, it has not been attempted for the problem at hand. However, Hartle demonstrated the principal feasibility of representing the information content of files, and the projection of the content onto the encoded symbol structures, in formal models that enable formal reasoning [6, 7]. This could be combined beneficially with approaches such as the typed object model proposed earlier [31].

How have been comparable problems be addressed in other disciplines?

The problem of quality-assuring products in industrial production uses an exact specification, calibrated measurement routines and devices, and approved model products that new products are compared with. In astronomy, a common problem is presented by turbulences in the atmosphere that distort light and thus blur the picture received on our planet's surface. To address this, adaptive optics use a reference light source to objectively measure distortions generated by the atmosphere and correct these distortions using a deformable mirror. Using a real star as a light source has severe limitations on the scope of applying the technique, since the ground truth is essentially unknown. Instead, a reference light source is generated, e.g. by a laser illuminating sodium gas in the mesosphere in about 15-25km height. By comparing the obtained distorted result with the known ground truth of the actual light source, the distortion function can be calculated. Using this reference source to correct atmospheric distortions yields astonishing improvements [18].

In information theory, Shannon introduced approximations of English language with the goal of creating a message source that has the same statistical properties, on various orders of approximation, as a message from a "real" source [22]. He thus created a model capable of producing messages that correspond closely enough to "real" messages so that the communication problem essentially becomes equivalent to transmitting "real" messages. Corresponding to this, what is ultimately necessary is a mathematical model of digital objects that is capable of producing objects that correspond closely to "actual" objects, closely enough so that the problem of "understanding" (accessing, preserving) them becomes essentially equivalent to understanding, accessing and preserving "actual" objects.

In current information management and preservation, however, the approach so far has been to apply algorithmic extraction in the form of static analysis on real-world objects with essentially unknown characteristics. This creates a circular dependency, since the algorithms have not been verified in large-scale experiments on real content. Moreover, the few systematic experimentation initiatives that try to establish a baseline of measures are not publicly available, since the content they are based on cannot be shared by the respective organisations [15].

The fundamental problem is one of angle: We need to approach the analysis of digital content from a different side and construct baseline benchmarks with known ground truth from bottom-up. This can then be complemented by a perceptual validation from the user side that is correlated to objectively obtained measures.

Instead of characterising objects taken from real collections, the solid bootstrapping approach presented in this article relies on generating test data from a fact base, from

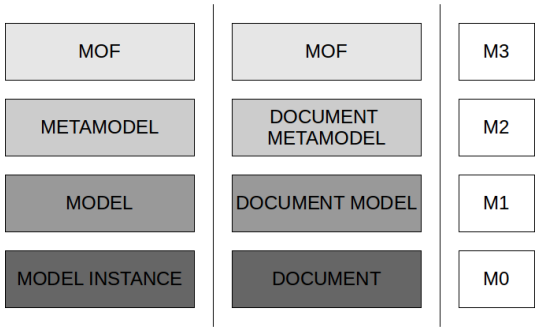


Figure 1: Meta-Object Facility layers

a content model with specified properties. This is supported by a novel application of model-driven engineering. To this end, the next section will introduce the basic background of model-driven engineering and how it can be applied to the design problem at hand.

3. MODEL-DRIVEN BENCHMARK DATA SETS GENERATION

Model driven engineering (MDE) addresses software complexity by shifting the focus of the engineer from algorithmic computing aspects to abstract representations of the knowledge and activities in relevant domains [21]. To this end, a *model* describes the level of reality which is needed for a certain purpose. This work uses model-driven engineering in a novel way where it adapts its methods and concepts so the final result is not an executable software product, but a test artefact.

3.1 Model Driven Engineering

The history of MDE starts with Computer Aided Software Engineering (CASE) but has evolved rapidly over the years. A leading approach in MDE is Model Driven Architecture (MDA)[8]. The key goals of MDA are to increase portability, interoperability and reusability through architectural separation of concerns, achieved by introducing three types of models:

- The *Computation Independent Model (CIM)* focuses on the conceptual perspective of a system's requirements in a certain context.
- The *Platform Independent Model (PIM)* focuses on the analytical specification perspective and describes the system and its operation independent of a specific target platform.
- The *Platform Specific Model (PSM)*, finally, focuses on the design and operation of a system specific to a particular platform.

The main value proposition lies in an increased ability to re-generate platform-specific artefacts based on the platform-independent model that can be managed and evolved.

MDA is based on standards such as UML and the Meta-Object Facility (MOF). MOF defines a 4-layer architecture for building metamodels and supporting core capabilities for model management, shown in Figure 1. The top level 3 provides support for defining metamodels at level M2. A *metamodel* is an abstraction of a domain in which concepts, relations among concepts and constraints that apply to those concepts and relations are defined. By creating a *metamodel*, a user defines a modelling language for building *models* which are defined at level M1. A prominent example of the relation between metamodels and models is UML, whose metamodel is specified in MOF. Once a model is defined, a running instance can be created on level 0. MOF provides a standardized platform for working with metamodels and models. Through *model transformations*, automatic translations of models based on one metamodel to another model based on the same or different metamodel are enabled. A *model transformation* is a set of mappings which translate one model into another. In MDA, this enables an automatic translation from PIM to PSM.

This work heavily relies on concepts defined by MDA and the 4-layer architecture defined by MOF. The right side of Figure 1 shows how the layered architecture is used in the case of generating documents. A key difference to mainstream MDA in this case is that the artefacts to be created are not necessarily software components and systems that should be maintained over time, but digital artefacts such as documents, images, databases and web pages. The intermediate artifacts used to create them can include software code, but the focus is not on improving interoperability and reusability *of that code*, but instead on gaining fine-grained control over the generation process of information objects to generate test data. In this case, the test data should ideally correspond closely enough to "actual" objects so that the problem of preserving it becomes essentially equivalent to preserving "actual" objects. This is of course an approximation problem, and it will be difficult to prove the exact degree of appromixation, just as it was with English language [22].

3.2 Conceptual framework

Figure 2 shows the main high-level framework of this work. There are four main aspects to the workflow: Analysis of real-world content, generation of fully annotated data sets that approximate real content, systematic automated evaluation, and open publication of the annotated data sets, artifacts, and experiment results.

Large-scale real-world content profiles of massive data collections are used to populate feature distribution models and domain-level as well as technical statistics and thus control the distribution of features and introduce fine-grained control over the feature space desired. This will be described in Section 3.3.

The resulting feature distributions inform the creation of platform-independent models for representing information objects such as page-based documents in an application-independent way. These can be automatically varied and diversified using predefined vocabulary elements in model mutation processes. Model transformation can project such model instances onto the space of platform-specific model representations (PSMs), from which specific artefacts can be generated. Section 3.4 presents examples of model transformations towards generating content.

Section 3.5 discusses the question of experimentation and measurement, while Section 4 will provide concrete example models and generated artefacts.

[8] `http://www.omg.org/mda/`

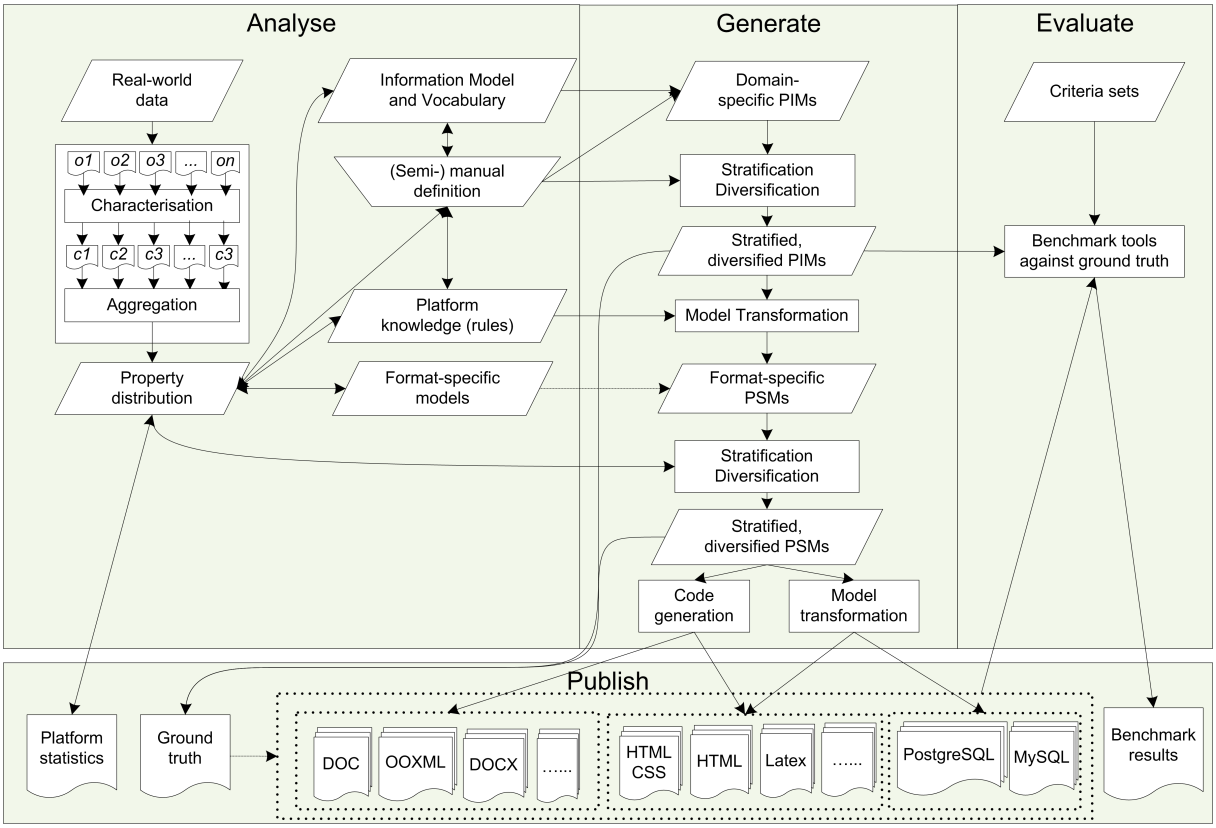


Figure 2: Model-driven benchmark generation framework

A key feature of this framework is the possibility to regenerate different representations of the same original information model with additional platforms when they become available, and to compare the direct transformation with representations created by conversion processes. This addresses some of the critical problems noted in the discussions of quality assurance [13].

3.3 Analyse

The step of large-scale analysis is supported by the content profiling system c3po[9], first presented in [17]. As shown in the left part of Figure 2, digital objects (*o1...on*) are first characterised, producing technical and descriptive metadata (*c1...cn*, each of which consists of a set of measures). These are then aggregated into a content profile, which can be analysed systematically.

The system analyses and aggregates the technical metadata generated by FITS on a scale-out map-reduce based platform (MongoDB)[10] with currently about 100.000 objects per minute per node. It supports multiple heuristics for selecting representative sample objects. This means it can be feasibly employed on a cluster to analyse very large collections. It is currently being tested on analysing almost half a billion web resources collected over a ten year period[11] in a single profile.

Aggregated statistics can be produced on any of the (potentially sparse) measures collected. Figure 3 shows exemplary property distributions for the number of documents with varying page counts (top) and the log-scaled number of web pages for occurence frequencies of the `<table>` tag.

It is beyond the scope of this paper to discuss the features themselves in detail. The key observation, however, is that we can measure aggregate statistical feature distributions on the content level (such as page counts) as well as on the technical level (such as tags and combinations of tags frequently used on web pages), and hence model the aggregate target distribution appriopriately. The hypothesis is that the complex interplay of such features is what leads to issues of understanding, rendering and hence preservation. Varying these feature sets in a controlled way, combined with auomated testing, should hence be a powerful tool for systematic assessment and improvement.

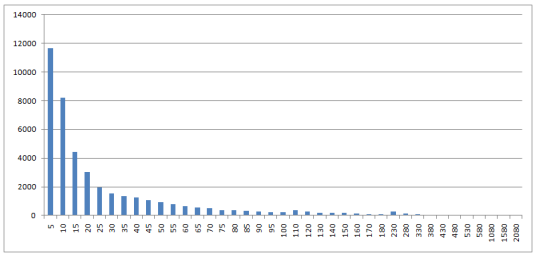
3.4 Generate: Models and transformations

3.4.1 Platform Independent Models

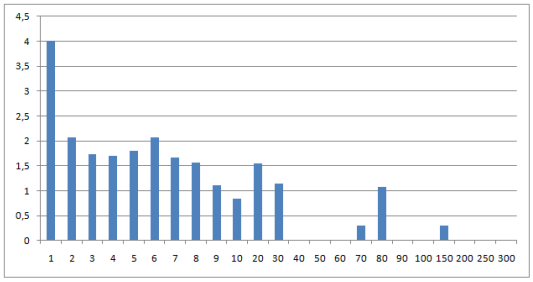
The starting point of generating annotated sets of content artefacts lies with a domain-specific PIM. This will typically correspond to a class such as *page-based documents* or *vector shapes*. Such PIMs are currently defined manually, but could in the future be derived from genre models.

A crucial question is the platform-independent representation of features that occur across different environments. Consider a grid structure of elements found on a page, such as items in a typical bill sent via email to an eCommerce customer. This grid structure can be represented in a platform-

[9] `http://http://ifs.tuwien.ac.at/imp/c3po`
[10] `http://www.mongodb.org/`
[11] `http://www.openplanetsfoundation.org/blogs/2013-01-09-year-fits`



(a) Page count distribution on 45.000 PDF documents



(b) Log-scaled `<table>` tag occurence in 82.000 web pages

Figure 3: Sample property distributions

independent model, but in any specific instance, it can be represented by a number of different constructs – a Word table, an Excel table embedded in a Word document as an OLE object (with or without formulas to calculate sums), tabbed text elements directly included in Word, a picture generated by an e-billing application, an HTML table construct, or an HTML/CSS construct for creating table structures, to name a few.

Within each of these feature spaces, there are a number of ways that elements can be grouped differently to achieve the same overall result. For example, the width of columns in a grid can be determined in a number of different ways such as percentage, automated layout, and fixed width settings. Real-world grids vary in dimensions such as size that are largely (but not fully) representation-invariant. They also exhibit representation-specific variation across the feature space of the continuously evolving platforms that are used to create and transmit them.

The independent representation can be projected onto specific instantations using different symbol structures that cause the same *performance* to be created. This can be fine-tuned to represent the variety that is encountered in real-world collections, but it can also be tuned to represent boundary conditions for testing the behaviour of analysis and rendering environments.

3.4.2 Metamodels and transformation steps

The idea behind the Platform Independent Meta-Model is to model building blocks and their relationships without a concern about platform specific implementation features. Figure 4 shows part of a platform independent metamodel for web pages. The model can contain several *Document* elements, each of which consists of zero or more *AbstractElements*. Apart from *ContainerElements*, the diagram shows three types of concrete *ContentElements*: *Media*, *Text*, and a *Grid*. A *Text* is described as a composition of *Sequence-*

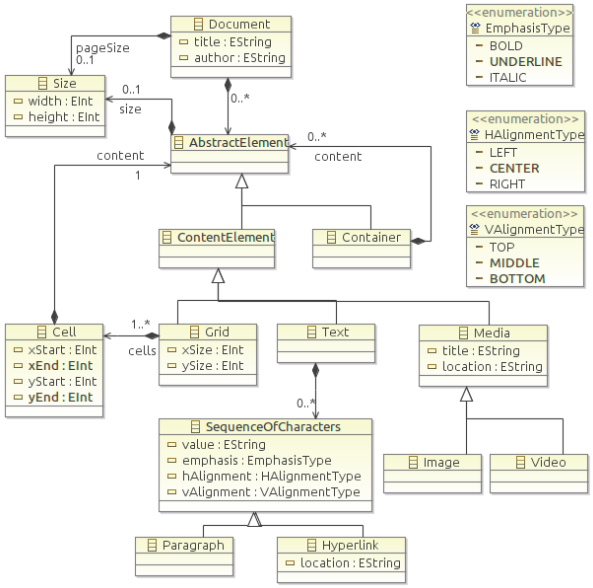


Figure 4: PIM metamodel for web documents

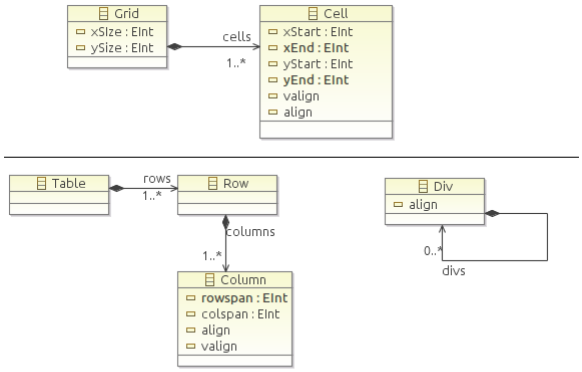


Figure 5: Multiple PSM variants of `Grid`

OfCharacters, including simple formatting options. A *Grid* element contains an array of *Cells* with coordinates. On the right side, media elements include images and video. While this simple model will hardly suffice to approximate real-world content, it serves to demonstrate the projection from PIM to platform-specific representations.

Figure 5 shows a fragment of Figure 4 above and two corresponding PSM fragments below. In this example of the *Grid*, possible representations include the html construct `<table>` and a set of `<div>` tags with appropriate hierarchical groupings. Both of these representations occur in real web pages with varying probability, and each needs to be decoded and rendered appropriately by a viewing application such as a browser.

Further model transformation and code generation elements allow the construction of born-digital information objects created directly within their native environments. Generation of content from the platform-specific models can employ two principal means: For certain well-defined content

types, objects can directly be created from the PSM. This requires a descriptive specification of the projection which can be formally expressed and verified. For content types with more particular variety, where the semantics are often shared between the objects and the creation environment, code generation will be employed that automates the native environments to create the objects directly within these environments. For example, Office automation languages can be used to control applications such as Microsoft Word® in writing, formatting, and exporting page-based and web documents in much the same way that human editors do. This effectively creates a simulation of a typical content production and editing process, and it allows the framework to include code that observes the state of the application at the time of creating the artefact and documents it to create additional traceability between the representations.

3.4.3 Diversification

Diversification processes can be used to create multiple instances with controlled variation of properties on both the intellectual domain and the technical implementation level. This diversification is informed by the statistics obtained from real-world data analysis.

3.4.4 Ground truth annotation

At each level of transformation, a set of properties can be documented to feed into the ground truth specification. While the ground truth of the PIM refers can connect models such as the one shown in Figure 4 to significant property specifications defined by domain experts [2], the PSM-specific documentation should also include the description of the features used to represent the PIM elements. Finally, if active code is used to automate content creation environments, this code can actively observe the state *inside* the application to document the original performance of creating the document.

3.5 Evaluate

Experimentation and automated measures have been a topic of intense research and development in DP. To evaluate precision and recall of characterisation tools, we need to define the property sets that describe the measures to be taken. This can be derived from decision criteria collected in preservation planning [2]. Current work is developing an ontology to represent these criteria and enable characterisation tools to declare which properties they measure. This enables direct integration through experimentation platforms such as the experimentation environment *Taverna*[12] and the workflow sharing platforms *myExperiment*[13].

To evaluate the data set generation process itself on a technical level, the expressiveness of the model-driven benchmark generation approach can be quantified with respect to relevant properties and features of specific content types as described in [2]. The coverage of these significant features of various content types can be calculated in a straightforward way. Similarly, the coverage of the content generation framework in terms of recreating the diversity of large-scale real-world content profiles can be assessed and quantified.

4. RESULTS AND DISCUSSION

[12] http://taverna.org.uk/
[13] http://myexperiment.org/

```
mapping PIM::Cell :: Cell2Cell() : PIM::Cell
{var horizontalAlignmentAsString :=
  GetValueFromDistribution("HorizontalAligmentTableCell",
        GetCellScenario(self)).repr();
 switch
 {
  case(horizontalAlignmentAsString = "Left") {
        horizontalAlignment := HorizontalAlignment::Left };
  case(horizontalAlignmentAsString = "Center") {
        horizontalAlignment := HorizontalAlignment::Center };
  case(horizontalAlignmentAsString = "Right") {
        horizontalAlignment := HorizontalAlignment::Right };
 };
  content += self.content.CellContent2CellContent();
}  ...
```

Figure 6: Example projection using QVT

4.1 Proof-of-concept domains and platforms

In order to evaluate the proposed approach, we have developed several proof-of-concept workflows for generating documents and databases. We will focus the discussion on two document-oriented experiments: One generated page-based documents and uses Macro-code generation and a runtime Office® environment to create Word files, while the other relies on XML and XSLT technologies to produce a variety of web pages. The purpose of the experiments was to validate the feasibility of the proposed framework, evaluate the applicability of different modelling and transformation technologies to the dataset generation problem, and analyse in more depth the challenges and key questions arising in the model transformation workflow.

4.1.1 Page-based documents

The creation of page-based documents relied on the Eclipse Modelling Framework (EMF)[14]. Transformation from PIM to PSM as well as diversification to a given property distribution is implemented using the standard Query/View/-Transformation [16]. Figure 6 shows a QVT fragment which diversifies the *alignment* property of a cell in a grid structure according to a stochastic property distribution similar to the one shown in Figure 3.

We used the statically typed template language Xpand[15] for generating VisualBasic macro code. The resulting macros were executed in Office 2007 and 2003. The features covered include tables, embedded OLE documents, embedded images, standard layout and fonts features, as well as page breaks, headers and footers. Figure 7 shows a code fragment from a generated artefact that creates a document, writes *Hello world*, and adds an embedded Excel sheet with a few cells containing numbers. The bottom lines of code query the state of the original application at the time of generating the artefact.

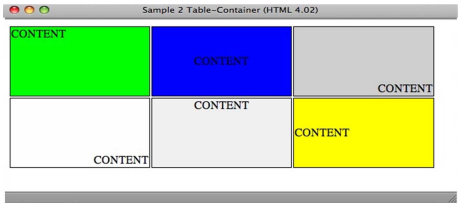
4.1.2 Hierarchical documents: web pages

For the creation of hiearchical documents, XML technologies were chosen for reasons of simplicity. An XML schema was defined for the PIM, and a set of transformation rules covered transformations to HTML 3.2, HTML 4.02, HTML 4 + CSS, and HTML 5. Features covered included tables, hyperlinks, sizes and alignment, font properties, colors, but also codecs and plugins. While this approach only relies on the availability of an XSLT processor, it does not directly support the expressive model transformation and code generation features delivered by EMF.

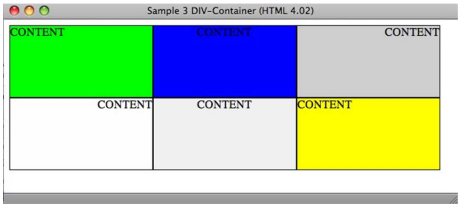
[14] http://www.eclipse.org/emf/
[15] http://wiki.eclipse.org/Xpand

```
Function Out_9_Word07(outputPath As String) As String()
    Documents.Add Template:="Normal", NewTemplate:=False,
        DocumentType:=0
    ...
                Selection.Font.Size = 16.0
                Selection.TypeText Text:="Hello"
                Selection.Font.Size = 9.0
                Selection.TypeText Text:="World"
        Selection.InlineShapes.AddOLEObject
                ClassType:="Excel.Sheet.12",
                LinkToFile:=False, DisplayAsIcon:=False
        Dim xlApp As Object
        Set xlApp = GetObject(, "Excel.Application")
        With xlApp.Application
          .Cells(1, 1).Select
          .ActiveCell.Font.Size = 10.0
          .ActiveCell.FormulaR1C1 = "Foo"
          .Cells(1, 2).Select
          .ActiveCell.Font.Size = 20.0
          .ActiveCell.FormulaR1C1 = "Bar"
          .ActiveCell.HorizontalAlignment = xlRight
        End With
        Set xlApp = Nothing
                Selection.Font.Size = 15.0
                Selection.TypeText Text:="Goodbye."
...
        ActiveDocument.Repaginate
        numberOfPages = ActiveDocument.
          BuiltInDocumentProperties(wdPropertyPages)
        numberOfWords = ActiveDocument.Range.
          ComputeStatistics(wdStatisticWords)
        numberOfCharacters = ActiveDocument.Range.
          ComputeStatistics(wdStatisticCharacters)
...
```

Figure 7: Generated macro code fragment



(a) Grid represented using `<table>`



(b) Grid represented using `<div>`

Figure 8: Partial mapping of PIM features in a PSM

As an alternative test that relied directly on a visual inspection of test results, we rendered the generated set of documents automatically using the cloud service *browsershots*[16]. This creates a visual snapshot of the artefacts using an entire array of environment configurations. Figure 8 shows an example of the above-mentioned *Grid* structure, projected to different HTML representations and rendered in a browser. The visual shows that the vertical alignment (a feature of the PIM) is not present in the rendering that uses `<div>` tags. The reason is that `<div>` tags cannot represent the vertical alignment feature, which leads to a partial mapping from PIM to PSM. This also points to a limitation of the simple XML/XSLT approach, which does not directly provide the sophisticated model verification facilities of EMF. However, the approach facilitated fast comparison of multiple renderings across browsers and has low technical requirements.

4.2 Discussion

While the scope of this paper prevents an in-depth discussion of the feature sets, technical intricacies of mappings, and detailed discussion of model transformation, it becomes clear that the basic approach of generating annotated test data is viable and enables the fine-grained control of test data. Combining this with large-scale data analysis creates a pwoerful framework to support the experimental evaluation of characterisation tools. This implies a number of clear benefits.

- **Open dataset.** All of the data sets generated by this framework can be freely published and redistributed without requiring careful verification of copyright, as is often the case with collected data [14].
- **Re-generation of content models across content platforms** is readily supported, which provides for a powerful extension mechanism to cover additional technology platforms and feature sets and can enable customized data set generation.
- **Full annotation of data sets with detailed ground truth**. The degree of detail that is provided as annotation with the data set can be fine-tuned to the degree desired. It should in all cases comprise the key attributes of PIM and PSM. In the case of automating a content creation environment to simulate a real creation process, even the actual state of the application can be observed and documented.
- **Extensible, reusable framework.** All structured knowledge produced in activities is passed on in fully specified models. All these models passed between activities can eventually be entry points interfacing with external systems.
- **Standard metrics** can be applied in a straightforward way to quantify the quality of tools, similar to established fields such as Information Retrieval. This can be a powerful market mechanism for both research and innovative practice, as can be seen on public competitions in IR such as TREC[17] [20].

Hence, the approach enables quantitative experimentation as a key instrument for scientific rigor and practical improvement in digital preservation. Baseline benchmark data sets will be created and used to validate existing methods and tools for content analysis against decision criteria sets in digital preservation. The results discussed above demonstrate the technical feasibility of the approach. However, before these benefits can be realised, several obstacles of the technical kind have to be overcome:

- The notion of *platform* as describing a technical platform such as *Mirosoft Word 2007* falls short of capturing the complex features that can be freely combined and mixed, even within a single (composite) object. Instead, we need to view the platform level as describing a *feature set*. This increases the complexity of the workflow generation.
- Much richer models will be required to achieve real-world approximation of documents, appeal to practitioners, and demonstrate that boundary conditions of

[16] `http://browsershots.org/`

[17] `http://trec.nist.gov/`

automated tools can be explored in a realistic, but controlled way.

- An experimentation platform for measures, including an ontology for metrics, is needed to systematically publish datasets and experiment results. This can build on existing Linked Data platforms such as LDS3 [24].

While the initial statistical analysis is relying on characterisation tools itself, it only relies on aggregate-level statistics, not individual object-level precision. Even if the errors are not evenly distributed and hence the distribution of features in fact incorrectly reported with a systemic bias, this does only affect the approximation accuracy of the feature distribution, not the precision of individually created items. Furthermore, as soon as these items are created, the systemic bias can be discovered and corrected. It is only through such a bootstrapping approach that the 'black box conundrum' can be tackled.

5. CONCLUSION AND OUTLOOK

It is clear that the DP field direly misses proper frameworks for benchmarking of the key information processing components, most importantly content characterisation. Going back to the dictionary, we find "*1: usually bench mark: a mark on a permanent object indicating elevation and serving as a reference in topographic surveys and tidal observations. 2 a: a point of reference from which measurements may be made. b: something that serves as a standard by which others may be measured or judged. c: a standardized problem or test that serves as a basis for evaluation or comparison (as of computer system performance)*"[18]. Key building blocks for benchmarking hence need to include

- A clear, unambiguous understanding of the processes that we want to measure,
- a clear set of attributes and indicators for taking measures,
- a well-defined value system for judging and assessing measures,
- solid hypotheses that can be tested and falsified,
- public, openly available data sets that can be shared and referenced,
- ground truth that annotates these data sets with useful and trusted measures corresponding to the attributes above, where trusted at least means that we know how reliable they are (something which is almost entirely absent in the data sets currently available), and
- a means for publication of benchmark results and all of the above elements.

This article discussed the state of art in such benchmarks in Digital Preservation and showed that a new approach is needed to successfully move research and development on content analysis and characterisation to a systematic, fundamentally solid approach. We outlined a conceptual framework for generating test data sets using Model Driven Engineering and demonstrated the technical feasibility of the approach.

Possessing a benchmark suite for information objects opens up new lines of R&D in digital preservation. It also enables commercial vendors, which have been reluctant to provide components specifically for digital preservation purposes, to leverage Return on Investment as a quantified benefit of improvement over existing solutions, something which currently is profoundly absent in digital preservation and which presents a fundamental inhibitor for innovation. As such, it can spur investment into specific products. Moreover, it will enable research to be much more focussed on essential key performance indicators. The principles of the framework can be extended easily for different content types. Quantitative experimentation will enable a deep understanding of the causal relationships of loss of authenticity, degrading, and digital decay. By empirically analysing which feature sets have which effects in which processes, we can enable simulation of digital decay over time. At the same time, establishing objective, quantitative performance indicators such as Precision and Recall for Digital Preservation will enable us to integrate on a systematic basis hitherto separated and isolated subdisciplines such as Information Retrieval and Digital Preservation.

One limitation that needs to be further explored is the question of malformed object artefacts. To investigate the behaviour of characterisation tools on objects that are *outside* of the technical specifications, i.e. the format constraints, we require data sets of objects that are carefully designed, and thoroughly annotated, to violate specific constraints and produce specific expected behaviour. These would represent meaningful test data sets that can be used to benchmark the robustness of existing preservation tools and processes and uncover the causes of typical errors. While this can in principle be supported by model-driven generation, it requires a model of these errors to be explicitly represented and addressed. This has not yet been attempted.

The medium-term objective is to showcase the generation of realistic approximations of common real-world information objects for at least page-based documents, web pages and databases, where the variation and combination of features can be controlled on a fine-grained basis and full ground truth is documented with the benchmark data. The expected benefits can be validated on a number of levels, starting with the achieved distribution of significant features over generated collections and the amount of errors in preservation processes that can be uncovered using the resulting benchmarks. Finally, we will use the benchmark collection to create baseline benchmark assessments of characterisation tools such as JHOVE2 and those delivered by projects such as SCAPE to measure precision and recall of relevant features for different formats. The generated benchmark data sets and their associated ground truth will be published with a royalty-free license, alongside evaluation results of tools common in the DP community.

The next steps correspondingly are the setup of a platform for experimentation and benchmark data set publication; the creation of large-scale platform statistics and longitudinal analysis based on 400 million web archive objects [19]; implementation of in-depth diversification and feature distributions; and the systematic testing of the functional correctness of commonly used tools.

[18] `http://www.merriam-webster.com/dictionary/benchmark`

[19] `http://www.openplanetsfoundation.org/blogs/2013-01-09-year-fits`

Acknowledgements

Part of this work was supported by the Vienna Science and Technology Fund (WWTF) through the project *BenchmarkDP* (ICT12-046), and by the European Union in the 7th Framework Program, IST, through the *SCAPE* project, Contract 270137. The authors wish to thank Harald Mezensky, Clemens Sauerwein and Florian Stäuble for carrying out some of the work that led to this article.

6. REFERENCES

[1] S. Abrams, S. Morrissey, and T. Cramer. 'What? So What?' The Next-Generation JHOVE2 Architecture for Format-Aware Characterization. In *Proc. IPRES*, 2008.

[2] C. Becker and A. Rauber. Decision criteria in digital preservation: What to measure and how. *JASIST*, 62(6):1009–1028, June 2011.

[3] C. Becker and A. Rauber. Preservation decisions: Terms and Conditions Apply. Challenges, Misperceptions and Lessons Learned in Preservation Planning. In *Proc. JCDL 2011*, June 2011.

[4] A. Dappert and A. Farquhar. Significance is in the eye of the. stakeholder. In M. Agosti, J. Borbinha, S. Kapidakis, C. Papatheodorou, and G. Tsakonas, editors, *Proc. ECDL*, volume 5714 of *LNCS*, pages 39–50. Springer, September 2009.

[5] M. Guttenbrunner and A. Rauber. A measurement framework for evaluating emulators for digital preservation. *ACM Transactions on Information Systems (TOIS)*, 30(2), 3 2012.

[6] M. Hartle, A. Botchak, D. Schumann, and M. Mühlhäuser. A Logic-based Approach to the Formal Description of Data Formats. In *Proc. IPRES*, pages 292–299, London, United Kingdom, September 2008. The British Library.

[7] M. Hartle, F.-D. Möller, S. Travar, B. Kröger, and M. Mühlhäuser. Using Bitstream Segment Graphs for Complete Data Format Instance Description. In *Proc. ICSOFT*, pages 198–205, Porto, Portugal, 2008.

[8] P. Heroux, E. Barbu, S. Adam, and E. Trupin. Automatic ground-truth generation for document image analysis and understanding. In *Proc. ICDAR*, ICDAR '07, pages 476–480, Washington, DC, USA, 2007. IEEE Computer Society.

[9] H. Heslop, S. Davis, and A. Wilson. An approach to the preservation of digital records. Green paper, National Archives of Australia, 2002.

[10] M. Hutchins. Testing software tools of potential interest for digital preservation activities at the national library of australia. Technical report, National Library of Australia, 2012.

[11] ISO/IEC. *Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models (ISO/IEC 25010)*. International Standards Organisation, 2011.

[12] A. N. Jackson. Formats over time: Exploring UK web history. In *Proc. IPRES*, 2012.

[13] N. Milic-Frayling. Digital object characterization: Document conversion and quality assurance. In *Automation in Digital Preservation: Dagstuhl Seminar Proceedings 10291*, Germany, 2010. `http://drops.dagstuhl.de/opus/volltexte/2010/2901`.

[14] R. Neumayer, C. Becker, T. Lidy, A. Rauber, E. Nicchiarelli, M. Thaller, M. Day, H. Hofman, and S. Ross. Development of an open testbed digital object corpus. DELOS Digital Preservation Cluster, Task 6.9, March 2007.

[15] 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 *Proc. of the 2nd DELOS Conference on Digital Libraries*, 2007.

[16] Object Management Group. Meta Object Facility (MOF) 2.0 Query/View/Transformation Specification, 2008. `http://www.omg.org/spec/QVT/1.0/PDF`.

[17] P. Petrov and C. Becker. Large-scale content profiling for preservation analysis. In *Proc. IPRES*, 2012.

[18] F. Roddier. *Adaptive Optics in Astronomy*. Cambridge University Press, 1999.

[19] J. Rothenberg. Ensuring the longevity of digital documents. *Scientific American*, 272, 1995.

[20] B. R. Rowe, D. W. Wood, A. N. Link, and D. A. Simoni. Economic Impact Assessment of NIST's Text REtrieval Conference (TREC) Program. RTI International, July 2010.

[21] D. Schmidt. Model-driven engineering. *IEEE Computer*, 39(2):25, 2006.

[22] C. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423 and and 623–656, July and October 1948.

[23] T. Strecker, J. v. Beusekom, S. Albayrak, and T. M. Breuel. Automated ground truth data generation for newspaper document images. In *Proc. ICDAR*, ICDAR '09, pages 1275–1279, Washington, DC, USA, 2009. IEEE Computer Society.

[24] D. Tarrant and L. Carr. LDS3: Applying digital preservation principals to linked data systems. In *Proc. IPRES*, 2012.

[25] M. Thaller, editor. *The eXtensible Characterisation Languages – XCL*. Verlag Dr Kovac, 2009.

[26] The SCAPE project. D9.1 characterisation technology, release 1 and release report. Technical report, SCAPE, 2012.

[27] The SCAPE project. Evaluation of characterisation tools. Technical report, SCAPE, 2012.

[28] D. Tkaczyk, A. Czeczko, K. Rusek, L. Bolikowski, and R. Bogacewicz. Grotoap: ground truth for open access publications. In *Proc. JCDL*, JCDL '12, pages 381–382, New York, NY, USA, 2012. ACM.

[29] L. Todoran, M. Worring, and M. Smeulders. The uva color document dataset. *Int. J. Doc. Anal. Recognit.*, 7(4):228–240, Sept. 2005.

[30] C. Webb, D. Pearson, and P. Koerbin. 'Oh, you wanted us to preserve that?!' Statements of Preservation Intent for the National Library of Australia's Digital Collections. *D-Lib Magazine*, 19(1/2), January/February 2013.

[31] J. M. Wing and J. Ockerbloom. Respectful type converters. *IEEE TSE*, 26(7):579–593, 2000.