Journal of Universal Computer Science, vol. 14, no. 18 (2008), 2936-2952
submitted: 1/3/08, accepted: 29/9/08, appeared: 1/10/08 © J.UCS

Systematic Characterisation of Objects in Digital Preservation: The eXtensible Characterisation Languages

Christoph Becker, Andreas Rauber
(Vienna University of Technology, Austria
{becker,rauber}@ifs.tuwien.ac.at)

Volker Heydegger, Jan Schnasse, Manfred Thaller
(University of Cologne, Germany
{herrmanv,jan.schnasse,manfred.thaller}@uni-koeln.de)

Abstract: During the last decades, digital objects have become the primary medium to create, shape, and exchange information. However, in contrast to analog objects such as books that directly represent their content, digital objects are not usable without a corresponding technical environment. The fast changes in these environments and in formats and technologies mean that digital documents have a short lifespan before they become obsolete. Digital preservation, i.e. actions to ensure longevity of digital information, thus has become a pressing challenge. The dominant strategies prevailing today are migration and emulation; for each strategy, different tools are available. When converting an object to a different representation, a validation of the content is needed to verify that the transformed objects are still authentically representing the same intellectual content. This validation so far is largely done manually, which is infeasible for large collections.

Preservation planning supports decision makers in reaching accountable decisions by evaluating potential strategies against well-defined requirements. Especially the evaluation of different migration tools for digital preservation has to rely on validating the converted objects and thus on an analysis of the logical structure and the content of documents. Existing approaches for characterising and describing objects do not attempt to fully extract the informational content of digital objects and thus are not suffficient for an in-depth validation of transformed content.

This paper describes the eXtensible Characterisation Languages (XCL) that support the automatic validation of document conversions and the evaluation of migration quality by hierarchically decomposing a document and representing documents from different sources in an abstract XML language. The description language XCDL provides an abstract representation of digital content in XML, while the extraction language XCEL allows an extraction engine to create such an abstract description by mapping file format structures to XCDL concepts.

We present the context of the development of these languages and tools and describe the overall concept and features of the languages. We further give examples and show how the languages can be applied to the evaluation of digital preservation solutions in the context of preservation planning.

Key Words: Digital Libraries, Digital Preservation, Preservation Planning, Content Characterisation, Migration, XML languages, Evaluation, File formats, File conversion

Category: H.3.7 Information Storage and Retrieval – Digital Libraries; I.7 Document and Text Processing

1 Introduction

The last decades have made digital objects the primary medium to create, shape, and exchange information. An increasing part of our cultural and scientific heritage is being created and maintained in digital form; digital content is at the heart of today's economy, and its ubiquity is increasingly shaping private lives.

The ever-growing complexity and heterogeneity of digital file formats together with rapid changes in underlying technologies have posed extreme challenges to the longevity of information. So far, digital objects are inherently ephemeral. Memory institutions such as national libraries and archives were amongst the first to approach the problem of ensuring long-term access to digital objects when the original software or hardware to interpret them correctly becomes unavailable [UNESCO, 2003].

Digital preservation denotes the efforts to preserve digital objects for a given purpose over long periods of time. The urgency of digital preservation has recently been reemphasised by the results of a survey among archiving professionals [The 100 Year Archive Task Force, 2007]. In the last years, numerous research initiatives have started around the world that aim at mastering this challenge.

The two strategies generally considered to prevail today are migration[Testbed, 2001; Mellor et al., 2002] and emulation[Rothenberg, 1999; van der Hoeven and van Wijngaarden, 2005]. Migration, the conversion of a digital object to another representation, is the most widely applied solution for standard object types such as electronic documents or images. The critical problem generally is how to ensure consistency and authenticity and preserve all the essential features and the conceptual characteristics of the original object whilst transforming its logical representation. Lawrence et. al. presented different kinds of risks for a migration project [Lawrence et al., 2000]. While migration operates on the objects and transforms them to more stable or more widely adopted representations, emulation operates on the environment of an object, trying to simulate the original environment that the object needs.

Consider a large collection of documents written in an old version of Word several years ago. This application does not run on modern operating systems. One could try to emulate the original operating system; but the emulation software still depends on current hardware, and emulating the hardware might be even harder. On the other hand, one could also migrate the documents to a current version of Word, or to PDF. While this would lose the original look-and-feel of the application, it would probably preserve the content and layout. With PDF, one would exchange the tool support provided by text processing software, as well as metadata such as an edit history, for a file format that is generally considered to be quite stable.

An important part of ongoing efforts in many large international projects is the outreach to vendors for advocating document engineering technologies for sustainable documents. The effects can be seen in standards such as PDF/A [ISO19005, 2004] or the Open Document Format (ODF) [ISO, 2006], or MPEG-7[ISO, 2002]. However, many objects exist and many more are created every day that face the threats of obsolescence. Hence, ex-post actions for preserving access to content are necessary. Performing actions on objects always risks damaging the content; but preserving authentic records also means being able to prove authenticity[ISSN 1082-9873, 2000; Rothenberg and Bikson, 1999].

Various migration tools are available for standard file formats such as office documents; the picture is less positive for more exotic and complex compound objects. However, even within migration tools for office documents, variation regarding the quality of conversion is very high. Some tools fail to preserve the proper layout of tables contained in a document; others miss footnotes or hyperlinks. Finding out which information has been lost during a conversion, and if this loss threatens the value of the object for a given purpose, is a very time-consuming task. Some losses might be acceptable, while others threaten the authenticity of documents. For example, if migrating the collection of Word documents mentioned above results in a loss of page breaks, this might be irrelevant if the textual content is the only thing of interest. However, if there are page references in the text, this loss might be unacceptable.

The complex situations and requirements that need to be considered when deciding which solution is best suited for a given collection of objects mean that this decision is a complex task. Preservation planning aids in the decision making process by evaluating available solutions against clearly defined and measurable criteria. This evaluation needs verification and comparison of documents and objects before and after migration to be able to judge migration quality in terms of defined requirements. It thus has to rely on an analysis of the logical structure of documents that is able to decompose documents and describe their content in an abstract form, independent of the file format. Especially considering migration actions working on large numbers of objects, it is essential to validate the authenticity of transformed objects automatically. When migrating a million documents from ODF to PDF/A, validation of these objects can not be done manually.

This paper provides an extended presentation of the experiments described in [Becker et al., 2008b]. It presents the eXtensible Characterisation Languages (XCL) that support the automatic validation of document conversions and the evaluation of conversion quality by hierarchically decomposing documents from different sources and representing them in an abstract XML language. We outline the basic concepts underlying the languages. We then describe the main architecture and features of XCL and discuss its application in the context of digital preservation.

The remainder of this paper is structured as follows. The next section outlines related work in the area of document engineering, digital preservation and the usage of XML in document conversion and extraction. Section 3 presents the Extensible Characterisation Languages and their usage within the context of Preservation Planning. Section 4 draws conclusions and points out directions for future work.

2 Related work

Digital preservation is a pressing matter – large parts of our cultural, scientific, and artistic heritage are exposed to the risks of obsolescence[UNESCO, 2003]. The rising awareness of the urgency to deal with the obsolescence that digital material is facing has led to a number of research initiatives over the last decade. A large part of the discourse has focused on discussing the dominant strategies, migration and emulation. Lawrence et. al.[Lawrence et al., 2000] analyzed risks of migration actions. While emulation is in principle widely applicable, the complexities and costs associated with it form a major obstacle to its wider adoption. Bearman strongly argues against the usage of emulation in [Bearman, 1999]. Rothenberg as one of the main proponents of emulation in digital preservation calls for encapsulation techniques to support emulation[Rothenberg, 1999]. Encapsulation as a complementary strategy packages the object to be preserved together with instructions on how it can be interpreted. Often the encapsulation layer is expressed in XML, which has a stronghold in digital preservation[Digital Preservation Testbed Project, 2002]. It is used not only for encapsulation, but also as a target file format for migration[Ramalho et al., 2007; Brandl and Keller-Marxer, 2007] or as a metadata container such as in PREMIS. [see 1]

In the discussion of file formats for long-term preservation it has recently been increasingly understood that the simple decision to use 'PDF' or 'TIFF' is highly problematic, as in both cases it is possible to create either files with very high as well as very low preservation value. In the case of extremely rich formats like PDF, this has led to the definition of subsets of the rules comprising the format which are considered safe for preservation, and furthermore to a tendency to identify informal 'subformats' of file formats[The Library of Congress, 2007; ISO19005, 2004].

At the heart of a preservation endeavour lies preservation planning. It is a core entity in the Reference Model for an Open Archival Information System, the OAIS model, which is a widely used model for archives[Consultative Committee for Space Data Systems, 2002]. It is also a core part of the EU project 'Preservation and Long-Term Access via Networked Services' (PLANETS) [see 2] which is creating a distributed service oriented architecture for digital preservation.

[1] `http://www.oclc.org/research/projects/pmwg/`
[2] `http://www.planets-project.eu`

Strodl et. al.[Strodl et al., 2007] present the PLANETS preservation planning methodology that aids in reaching well-founded decisions. Becker et. al. describe the planning tool Plato actively supporting and automating the workflow.[Becker et al., 2008a]

The procedure defines measurable requirements for preservation strategies in a hierarchical form and evaluates them in a standardised testbed setting. Similarly, Ferreira [Ferreira et al., 2007] presents a Service Oriented Architecture for recommending and performing format migrations based on pre-specified requirements. However, so far most of the evaluation of tools against these requirements has to be done manually. For example, to evaluate if the layout of a document has been preserved during migration, a human has to look at the files and compare them with each other. This is not feasible for large collections; automated services have to be integrated that characterise content to support this evaluation.

A number of tools and services have been developed that perform content characterisation specifically for digital preservation. The National Library of New Zealand Metadata Extraction Tool [see 3] extracts preservation metadata for various input file formats. Harvard University Library's tool JHove [see 4] enables the identification and characterisation of digital objects. However, both tools only extract metadata such as the presence of specific file format features in a document; they do not describe the content of a document.

Some solutions exist for transforming, matching, and comparing structured documents. Díaz describes the usage of XML for handling the conversion problems that arise in the exchange of business documents between organisations[Díaz et al., 2002]. Canfield presents an algorithm for approximate XML document matching in [Canfield and Xing, 2005]. In the area of grid computing, the Global Grid Forum Data Format Description Language Working Group has been working on a language called DFDL[Beckerle and Westhead, 2004] which extends XML Schema. The aim is to describe *the structure of binary and character encoded (ASCII/Unicode) files and data streams so that their format, structure, and metadata can be exposed* [see 5]. Thus the DFDL creates a mapping between formatted files and corresponding XML representations. The PADS language, on the other hand, is a domain-specific language based on C-structures that aims at performance-oriented processing of large, simple structured files[K.Fisher and Gruber, 2005].

[3] `http://meta-extractor.sourceforge.net/`
[4] `http://hul.harvard.edu/jhove`
[5] `http://forge.ggf.org/sf/projects/dfdl-wg`

3 The extensible characterisation languages

3.1 Introduction

The eXtensible characterisation languages consist of two languages: the description language XCDL and the exctraction language XCEL. The specifications for both languages are publicly available [see 6].

The Extensible Characterisation Description language (XCDL) allows the representation of characteristics extracted from files. The definition of characteristics, however, is taken in the broadest possible way: So, conceptually, an XCDL representation of the information contained within a file can be a complete interpretation of all the information contained in that file.

Converting any number of documents from one format into another, i.e. transforming the actual representation of its content, inevitably raises the issue of preserving authenticity. Moreover, to confidently choose between alternative target formats and tools, one has to evaluate their suitability in a given context. This leads to the following underlying questions.

1. Which information contained in the old format is also contained in the new format?
2. Which information relevant to the usage of the content of the old format is contained in the new format?
3. Is the conversion process *a(old,new)* better then *b(old,new)*, i.e. does it preserve more of the relevant information contained within the object?

Comparing information in two different file formats implies the following requisites.

1. An abstract way of expressing the information in a format-neutral model. This is henceforth called an *extensible characterisation definition language (XCDL)*.
2. A way of extracting information in specific file formats and describing it using XCDL. While it would be theoretically possible to create an extraction tool for every given file format, in practice this is not feasible. A better solution is to define a comprehensive *extensible characterisation extraction language (XCEL)* and implement an extractor component that is able to interpret this language.
3. Algorithms for comparing two XCDL descriptions for degrees of equality.

[6] `http://planetarium.hki.uni-koeln.de/public/XCL`

Such a language should be defined so generic that it supports the description of arbitrary file formats and thus the extraction of characteristics from *any* given file.

An XCDL document describes the content of a specific file with format type X, tagged in XML according to the XCDL language specifications, and is processible through an XCDL interpreter. An XCEL document describes what information can be extracted from any given file of format X, enabling an XCEL processor to extract this information and express it in XCDL. XCEL thus creates a mapping between the declarative description of the information in a physical file and its abstract interpretation outside of a format specification. Both XCDL and XCEL are meta-languages defined using XML Schema. In contrast to other applications of XML in digital preservation, XCL does not migrate digital objects as a whole to XML nor store exclusively preservation metadata; it transforms the entire content of an object into an abstract unified form. A key application we focus on is the comparison of different representations of the same object in order to validate migration within the preservation planning procedure.

The next sections will describe the basic architecture of both characterisation languages. We will then outline an example of how they can be applied in practice.

3.2 The extraction language XCEL

The eXtensible Characterisation Extraction Language (XCEL) is a file format description language for describing file structures to allow their parsing by generalized software. The main goal of the XCEL is therefore to provide all tools necessary for describing real-life file formats like PNG, TIFF, PDF or DOCX. [see 7] The XCEL is a declarative, descriptive, XML-based language that provides well defined mechanisms for extending certain parts of the language. As an Extraction Language the XCEL has some similarities to other domain specific languages for describing file formats.[Fisher et al., 2006] The Data Format Description Language (DFDL) has a number of common properties with the XCEL, there are significant differences, however. DFDL focuses on scientific data3, XCEL is primarily targeted at file formats as typically held in libraries and archives. The DFDL is implemented as an extension of XML-Schema, the XCEL is a completely new language the syntax of which can be described with the XML-Schema language. Other approaches like PADS are focusing the processing of simple but

[7] The PNG specification is available at `http://www.w3.org/TR/PNG/`.
The TIFF specification is available at `http://partners.adobe.com/public/developer/tiff/index.html`.
The PDF specification is available at `http://www.adobe.com/devnet/pdf/pdf_reference.html`.
The DOCX specification is available at `http://www.ecmainternational.org/news/TC45_current_work/TC45_available_docs.htm`

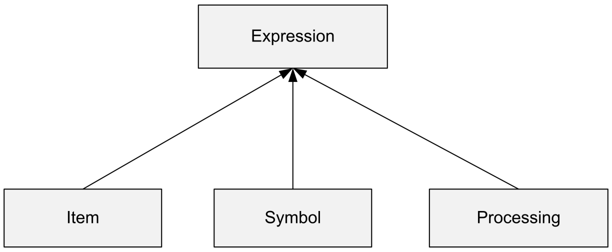

Figure 1: The structuring elements of XCEL

large scale data formats[K.Fisher and Gruber, 2005]. The distinct characteristic of XCEL, however, is, that it does not extract purely technical entities such as 'a 3 x 256 array of one byte numbers', but characteristics with a semantic meaning - 'a colour lookup table'.

An XCEL document comprises the following components.

1. **Preprocessing** information includes configuration tasks affecting the behaviour of the XCEL interpreter.
2. The **format description** is the core part defining the structure of an object.
3. **Templates** describe recurring structures such as number formats in ASCII based file formats.
4. **Postprocessing** instructions define actions to be performed on the result of the format processing.

Figure 1 shows the abstract relations of XCEL expressions. This structure – which seems to be simpler than the one proposed by the DFDL [see 8] – is based upon the assumption that any file format can be expressed as (a) a set of hierarchies of blocks of content, all of which can (b) be addressed from within but also outside of these hierarchies.

An XCEL format description starts with an *Item*, a container element that can have different content models, similar to the XML-Schema content models. A *Symbol* is an expression that adds a name or ID to a specific part of the byte stream. The *Processing* element models an expression that is used to call built-in methods for the extraction processor. This structure describes file formats in a tree where each branch describes one possible variation. It is the job of the XCEL processor to find out which branch matches to a given file.

Figure 2 shows the XCEL description of the IDAT chunk of a PNG[ISO, 2004] image. Every chunk in PNG consists of the consecutive parts `length`,

[8] `http://forge.ggf.org/sf/projects/dfdl-wg`

```
<!-- The complete IDAT chunk is expressed as one item that
     prescribes all its children to appear in the given order -->
<item xsi:type="structuringItem" identifier="IDAT" multiple="true"
optional="true">
    <symbol identifier="IDATLength" interpretation="uint32" length="4" />
    <symbol identifier="IDATChunkType" interpretation="ASCII"
            value="IDAT" />
    <!-- Set the length of the expression "IDATChunkData"
       to the value given by the expression "IDATLength"-->
    <processing type="pushXCEL" xcelRef="IDATChunkData">
        <processingMethod name="setLength">
            <param valueRef="IDATLength"/>
        </processingMethod>
    </processing>
    <symbol identifier="IDATChunkData" interpretation="uint8"
            name="normData"/>
    <symbol name="IDATCRC" length="4" />
</item>
```

Figure 2: XCEL description of a PNG chunk

`chunk type`, `chunk data` and `CRC`, where the `length` is a four byte unsigned integer that contains the length of the `chunk data` field, `chunk type` is a four byte ASCII keyword, `chunk data` is a field that can contain any data structure, and `CRC` contains a checksum.

The XCEL processing software ('Extractor') processes the binary files of given formats using the specific XCEL documents created for these formats. Currently we have prepared XCEL documents for various file formats, focusing on the image, text and audio data domain (e.g. TIFF, PNG, GIF, WAV, and PDF). The Extractor is conceived in such a way as to be able to process any additionally created XCEL document without modifications of its core implementation. Thus, to enlarge the spectrum of supported file formats one only has to write an XCEL document for that format.

3.3 The description language XCDL

The result of extracting content from a file using an XCEL document as input for an extractor is a description of the informational content of this file in the description language XCDL. A simple example is provided in Figure 3, which provides a part of an XCDL description of a text document containing the phrase 'An **important** word'.

A common characteristic of content models is a separation between primary information and properties of that information. Within the textual domain this separation consists e.g. in the difference between the string 'An important word' and the means by which we indicate that the single words in that string are expressed in specific fonts. The corresponding XCDL representation is provided below. The *normData* tag wraps the primary information in a context-free manner, removing or transforming all format-specific information as well as its specific representation. We call this kind of representation normalised data. Text encoding is translated into UTF8 by default. The fonts are described within the *property* tags. In this case we have a property describing the fonts used. For each different font a value set is created. The font name appears as a labelled value referring to exactly defined terms in the XCL properties ontology. The *dataRef* tags define positions within the normalised data by referencing a *propertySet* which indicates where the specific fonts are applied. The *propertySet* furthermore contains references to all related *valueSet* entries, thus creating a bi-directional mapping. This basic structure of separating data and associated properties is common for all underlying content models: In the case of images, e.g., the primary stream of bytes describing the pixels can have properties, which are applicable to an image as a whole (e.g. a gamma correction) or only to parts thereof, as e.g., an embedded explanatory text in the image.

For preservation purposes, the properties extracted from a file may include a category of properties which are not needed to model the content of the file. Consider, e.g., a file format which compresses a part of the data it contains. A proper XCL extractor, which extracts the content of the file and expresses it in XCDL, has to be able to apply that algorithm in order to decompress the content. Once this is done, the algorithm applied to the original file becomes irrelevant, as the content is simply the result of its application. For preservation purposes - basically tracking the history of a file and its authenticity - properties like 'originally compressed by algorithm X' can be expressed.

```
<object id="o1" >
  <normData type="text" id="nd1">An important word</normData>
  <property id="p94" source="raw" cat="descr" >
      <name id="id86" >pdfBaseFont</name>
      <valueSet id="i_i1_i49_i2_i3" >
          <labValue>
              <val>NimbusRomanNo9L-Regu</val>
              <type>string</type>
          </labValue>
          <dataRef ind="normSpecific" propertySetId="id_0" />
      </valueSet>
      <valueSet id="i_i1_i49_i2_i4" >
          <labValue>
              <val>NimbusRomNo9L-Medi</val>
              <type>string</type>
          </labValue>
          <dataRef ind="normSpecific" propertySetId="id_1" />
      </valueSet>
  </property>
  <property id="p106" source="raw" cat="descr" >
      <name id="id158" >fontSize</name>
      <valueSet id="i_i1_i70_i2" >
          <labValue>
              <val>12</val>
              <type>rational</type>
          </labValue>
          <dataRef ind="normSpecific" propertySetId="id_0" />
      </valueSet>
  </property>
  <propertySet id="id_0" >
      <valueSetRelations>
          <ref valueSetId="i_i1_i49_i2_i3" name="pdfBaseFont" />
          <ref valueSetId="i_i1_i70_i2" name="fontSize" />
      </valueSetRelations>
      <dataRef>
          <ref begin="0" end="1" id="nd1" />
          <ref begin="13" end="16" id="nd1" />
      </dataRef>
  </propertySet>
  <propertySet id="id_1" >
      <valueSetRelations>
          <ref valueSetId="i_i1_i49_i2_i4" name="pdfBaseFont" />
          <ref valueSetId="i_i1_i70_i2" name="fontSize" />
      </valueSetRelations>
      <dataRef>
          <ref begin="2" end="12" id="nd1" />
      </dataRef>
  </propertySet>
</object>
```

Figure 3: XCDL representation of primary information and corresponding properties, connected by property sets

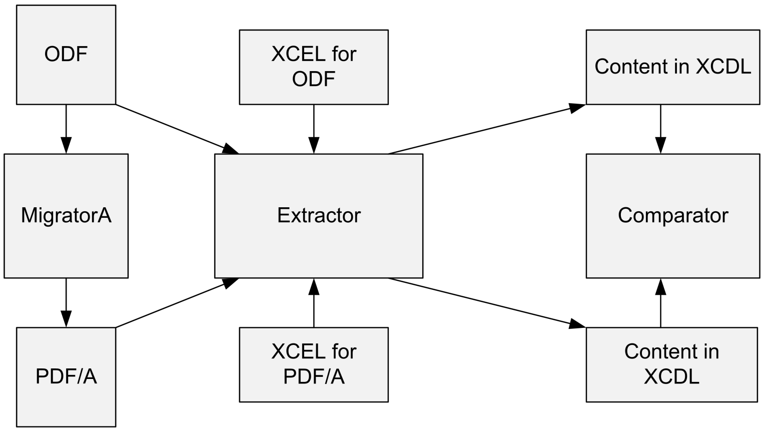

Figure 4: Using XCL to compare migrated documents

3.4 Comparing digital objects

The XCL languages have been designed with the primary goal of automating the validation of content in converted representations within the preservation planning procedure. Figure 4 shows a corresponding scenario for applying XCL in the context of format migration. After converting a document from ODF to PDF/A, the XCDL documents of the original and the transformed object can be compared using an interpretation software. A comparison tool ('Comparator') for XCDL documents is currently under development. Key objectives are the property-specific definition of metrics and their implementations as algorithms in order to identify degrees of equality between two XCDL documents. In its core functionality the comparator loads two XCDL documents, extracts the property sequences and compares them according to comparison metrics which are defined with respect to the types of the values in the value sets. In the example of Figure 3, the comparator looks up the defined metrics for property 'Fontname' and executes the comparison according to the metrics definition. This can be a simple binary comparison that checks the XCL ontology for the entries 'Times-Roman' and 'Times-Bold'. For other properties such as as possible deviation of font size, absolute or relative difference measures can be used.

This will provide a sophisticated means for widespread usage, e.g. to evaluate the quality of a migration process and thus support decisions in the preservation planning process.

To verify the approach we migrated a benchmark corpus of PNG files [see 9] to TIFF and compared the resulting XCDL documents. In contrast to other

[9] `http://www.schaik.com/pngsuite`

tools such as JHove or tiffInfo [see 10], XCL was able to extract file properties as well as the normalised content from all files. Comparing the *normData* with a tool revealed that conversion of images with specific characteristics was not working properly.

For evaluating preservation strategies, preservation planning activities define requirements that a solution has to meet. Often, a complete and extensive comparison is not needed. The comparator offers the possibility to select and weigh only a subset of properties, thus enabling users to regulate the relevance of different properties with respect to their specific needs. By mapping the content structures in XCDL as well as the results from the *Comparator* to the requirements, performance comparisons across different preservation strategies can be defined and recommendations for a solution can be given in an automated way.

3.5 Summary

The XCL languages presented in this paper provide a comprehensive and abstract model to describe and express properties of digital objects. The definition of an XCEL allows to describe these properties with a unique vocabulary. The implementation of an XCEL processor enables to extract object properties in an improved quality as XCDL documents. Digital objects described in XCDL can easily and effectively be processed by interpretation software. We develop such an interpretation tool to enhance and support digital preservation efforts.

4 Summary and Outlook

A range of tools exist today for migration between different document formats. These tools have very specific strengths and weaknesses. Some fail to preserve document layout properly, while others would lose content embedded in objects or fail to preserve structural relations. The evaluation of authenticity of transformed content is a complex task; so is the overall evaluation of suitability of these tools in a particular situation. Digital preservation is thus in need for automated characterisation services that support preservation planning in evaluating potential strategies. These services need an abstract means of describing a document's content in an interoperable, format-independent way.

When comparing the content of two files stored in two different formats, we have to distinguish between the abstract content and the way in which it is wrapped technically. On a very abstract level, this will for a long time be impossible: Whether an image of a hand-written note contains the same 'information' as a transcription of that note in UTF-8 is philosophically interesting,

[10] `http://remotesensing.org/libtiff/man/tiffinfo.1.html`

but scarcely decidable on an engineering level. In a more restricted way, a solution is possible if we express the content stored in different file formats in terms of an abstract model of that type of content.

The extensible characterisation extraction and definition languages presented in this paper are an important step towards this goal. The extraction language XCEL allows the extractor component to extract the content of any document provided in a format for which an XCEL specification exists. The content is described in the description language XCDL and can thus be compared to other documents in a straightforward way. This differentiates the XCL approach from the approach used by JHove and similar projects. The XCL does not attempt to extract a set of characteristics from a file, but it proposes to express the complete informational content of a file in a format independent model.

The DFDL language, on the other hand, concentrates on exact typing of data formats for interoperability on the grid. Consider a binary file with the content '`00000000 00100000`'. Using DFDL, it is possible to express that the file contains an unsigned 16 bit number in big endian encoding, i.e. 32. XCL is able to express that the file contains a 16 bit number in big endian encoding meaning *imageWidth=32*. Thus XCL also intends to describe the semantics of a file.

This paper outlined the basic architecture of the two characterisation languages, provided examples of how they can be applied in practice and discussed the potentials of the proposed approach in the context of digital preservation and preservation planning. Future work will be geared towards implementing automated verification and evaluation of different tools and integrating comparison services into the decision support software for preservation planning.

Acknowledgements

Part of this work was supported by the European Union in the 6th Framework Program, IST, through the PLANETS project, contract 033789.

References

[Bearman, 1999] Bearman, David: Reality and Chimeras in the Preservation of Electronic Records. In: *D-Lib Magazine* 5 (1999), April, No. 4

[Becker et al., 2008a] Becker, Christoph ; Kulovits, Hannes ; Rauber, Andreas ; Hofman, Hans: Plato: A Service Oriented Decision Support System for Preservation Planning. In: *Proceedings of the 8th ACM IEEE Joint Conference on Digital Libraries (JCDL'08)*, 2008

[Becker et al., 2008b] Becker, Christoph ; Rauber, Andreas ; Heydegger, Volker ; Schnasse, Jan ; Thaller, Manfred: A Generic XML Language for

Characterising Objects to Support Digital Preservation. In: *Proc. 23rd Annual ACM Symposium on Applied Computing (SAC'08)* Ed. 1. Fortaleza, Brazil : ACM, March 16-20 2008, pp. 402–406

[Beckerle and Westhead, 2004] Beckerle, M. ; Westhead, M.: GGF DFDL Primer / Global Grid Forum Data Format Description Language Working Group. 2004. – Technical Report

[Brandl and Keller-Marxer, 2007] Brandl, Stefan ; Keller-Marxer, Peter: Long-term Archiving of Relational Databases with Chronos. In: *First International Workshop on Database Preservation (PresDB'07).* Edinburgh, March 2007

[Canfield and Xing, 2005] Canfield, E. R. ; Xing, Guangming: Approximate XML document matching. In: *SAC '05: Proceedings of the 2005 ACM symposium on Applied computing.* New York, NY, USA : ACM Press, 2005, pp. 787–788. – ISBN 1-58113-964-0

[Consultative Committee for Space Data Systems, 2002] Consultative Committee for Space Data Systems: *Reference Model for an Open Archival Information System (OAIS).* CCSDS 650.0-B-1, 2002

[Díaz et al., 2002] Díaz, Luis M. ; Wüstner, Erik ; Buxmann, Peter: Inter-organizational document exchange: Facing the conversion problem with XML. In: *SAC '02: Proceedings of the 2002 ACM symposium on Applied computing.* New York, NY, USA : ACM Press, 2002, pp. 1043–1047. – ISBN 1-58113-445-2

[Digital Preservation Testbed Project, 2002] Digital Preservation Testbed Project: XML and Digital Preservation. 2002. – Technical Report. `http://www.digitaleduurzaamheid.nl/bibliotheek/docs/white-paper_xml-en.%pdf`

[Ferreira et al., 2007] Ferreira, Miguel ; Baptista, Ana A. ; Ramalho, Jose C.: An intelligent decision support system for digital preservation. In: *International Journal on Digital Libraries* 6 (2007), July, No. 4, pp. 295–304

[Fisher et al., 2006] Fisher, K. ; Mandelbaum, Y. ; Walker, D.: The next 700 Data Description Languages. In: *Conference record of the 33rd ACM SIGPLAN-SIGACT symposium on Principles of programming languages*, 2006, pp. 2–15

[van der Hoeven and van Wijngaarden, 2005] Hoeven, Jeffrey van der ; Wijngaarden, Hilde van: Modular emulation as a long-term preservation strategy for digital objects. In: *5th International Web Archiving Workshop (IWAW05)*, 2005

[ISO, 2002] ISO: *Information technology – Multimedia content description interface – Part 1: Systems (ISO/IEC 15938-1:2002).* International Standards Organization (Veranst.), 2002

[ISO, 2004] ISO: *Information technology – Computer graphics and image processing – Portable Network Graphics (PNG): Functional specification*

(ISO/IEC 15948:2004). International Standards Organization (Veranst.), 2004

[ISO, 2006] ISO: *Information technology – Open Document Format for Office Applications (ISO/IEC 26300:2006).* International Standards Organization (Veranst.), 2006

[ISO19005, 2004] ISO: *Document management - Electronic document file format for long-term preservation - Part 1: Use of PDF 1.4 (PDF/A) (ISO/CD 19005-1).* International Standards Organization (Veranst.), 2004

[ISSN 1082-9873, 2000] Gilliland-Swetland, A.J. ; Eppard, P.B.: Preserving the Authenticity of Contingent Digital Objects: The InterPARES Project. In: *D-Lib Magazine* 6 (2000), July-August, No. 7/8. – `http://www.dlib.org/dlib/july00/eppard/07eppard.html`

[K.Fisher and Gruber, 2005] K.Fisher ; Gruber, R.: A Domain-Specific Language for Processing Ad Hoc Data. In: *Proceedings of the 2005 ACM SIGPLAN conference on Programming language design and implementation*, 2005, pp. 295–304

[Lawrence et al., 2000] Lawrence, Gregory W. ; Kehoe, William R. ; Rieger, Oya Y. ; Walters, William H. ; Kenney, Anne R.: Risk Management of Digital Information: A File Format Investigation / Council on Library and Information Resources. URL `http://www.clir.org/pubs/abstract/pub93abst.html`, June 2000 (93). – CLIR Report

[Mellor et al., 2002] Mellor, P. ; Wheatley, P. ; Sergeant, D.M.: Migration on request, a practical technique for preservation. In: Agosti, M. (Eds.) ; Thanos, M.C. (Eds.): *Proceedings of the 6th European Conference on Digital Libraries (ECDL'02)*, Springer, 2002, pp. 516–526

[Ramalho et al., 2007] Ramalho, José C. ; Ferreira, Miguel ; Faria, Luís ; Castro, Rui: Relational Database Preservation through XML modelling. In: *Proceedings of Extreme Markup Languages 2007.* Montréal, Québec, August 2007

[Rothenberg, 1999] Rothenberg, J.: *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation.* Council on Library and Information Resources, January 1999. – `http://www.clir.org/pubs/reports/rothenberg/contents.html`

[Rothenberg and Bikson, 1999] Rothenberg, Jeff ; Bikson, Tora: Carrying Authentic, Understandable and Usable Digital Records Through Time / Report to the Dutch National Archives and Ministry of the Interior. The Hague, Netherlands, 1999. – Technical Report

[Strodl et al., 2007] Strodl, Stephan ; Becker, Christoph ; Neumayer, Robert ; Rauber, Andreas: How to Choose a Digital Preservation Strategy: Evaluating a Preservation Planning Procedure. In: *Proceedings of the 7th ACM IEEE Joint Conference on Digital Libraries (JCDL'07)*, June 2007, pp. 29–38

[Testbed, 2001] Testbed, Digital P.: Migration: Context and current status /

National Archives and Ministry of the Interior and Kingdom Relations. 2001. – White paper

[The 100 Year Archive Task Force, 2007] The 100 Year Archive Task Force: *The 100 Year Archive Requirements Survey.* http://www.snia.org/forums/dmf/programs/ltacsi/100_year/. 2007

[The Library of Congress, 2007] The Library of Congress: *Preferences in Summary for Textual Content.* Website. accessed August 2007. – http://www.digitalpreservation.gov/formats/content/text_preferences.sht%ml

[UNESCO, 2003] UNESCO: *UNESCO Charter on the Preservation of Digital Heritage.* Adopted at the 32nd session of the General Conference of UNESCO. October 17, 2003. – http://portal.unesco.org/ci/en/files/13367/10700115911Charter_en.pdf/Ch%arter_en.pdf