Project Number	IST-2006-033789
Project Title	Planets
Title of Deliverable	*Planets components for the extraction and evaluation of digital object properties*. (Part two of a three-part final report from the Digital Object Properties Working Group Report)
Deliverable Number	D23B
Contributing Sub-project and Work-package	PC/3
Deliverable Dissemination Level	External
Deliverable Nature	Report
Contractual Delivery Date	26th April 2010
Actual Delivery Date	26th May 2010
Author(s)	TNA, ONB, KBNL, UzK

Keyword list: Digital object properties, significant properties, characteristics, observational, extractable, ontology, PCR, preservation characterisation, preservation action, preservation planning.

EXECUTIVE SUMMARY

The rate of technological change and the dependency of digital objects on technology in order to be found, accessed, understood and utilised, means that there is a real risk of obsolescence for digital objects if they are not actively preserved.

Understanding, defining and assessing the individual properties of a digital object are important devices for informing decisions about which characteristics of that object should be preserved over time, in circumstances where it is not possible, for reasons such as cost, practicality or technical constraints, to preserve all the elements of that object.

This report sets out the components developed during the course of the PLANETS project to extract and evaluate significant properties, gives recommendations for future work, and in it's Appendix, illustrates through a use case how these components might be used.

This report is part of a three-part final report from the PLANETS Digital Object Properties Working Group. The three companion reports, which can be read in conjunction, are:

- *The concept of significant properties.* (PLANETS deliverable PC3 – D23A);
- *Planets components for the extraction and evaluation of digital object properties* (PLANETS deliverable PC3 – D23B (this report)); and
- *Specification of a Planets-wide Ontology of properties for digital preservation needs.* (PLANETS deliverable PC3 – D23C)

TABLE OF CONTENTS

1. Introduction

1.1 The purpose of this document

This report sets out the components developed during the course of the PLANETS project to extract and evaluate digital object properties, gives recommendations for future work, and in it's Appendix, illustrates through a use case how these components might be used.

This report is part of a three-part final report from the PLANETS Digital Object Properties Working Group. The three companion reports, which can be read in conjunction, are:

- *The concept of significant properties.* (PLANETS deliverable PC3 – D23A);
- *Planets components for the extraction and evaluation of digital object properties* (PLANETS deliverable PC3 – D23B (this report)); and
- *Specification of a Planets-wide Ontology of properties for digital preservation needs.* (PLANETS deliverable PC3 – D23C)

1.2 PLANETS and the Digital Object Properties Working Group

PLANETS (Preservation and Long-term Access through Networked Services), is a four-year project co-funded by the European Union, to address core digital preservation challenges. Started in 2006, the main aim of the project is to develop practical services and tools to help ensure long-term access to digital cultural and scientific assets. To this end, the project draws on the expertise of 16 project partners from national libraries and archives, leading research universities, and technology companies across Europe. Work within the project was divided between the six separate subprojects of Preservation Planning, Preservation Action, Preservation Characterisation, Testbed, Interoperability Framework and Dissemination and Training. Each of these was further divided into work packages.

Within the field of digital preservation, digital object properties play an important role in informing preservation planning, actions and characterisation. From the beginning of the PLANETS project, several of the different subprojects undertook work involving digital object properties, each with different approaches and focuses (see chapter 2 for more detail). In 2008, the Significant Properties Working Group was set up in order to assess the digital object properties needed to evaluate the behaviour of preservation tools within the PLANETS testbed, and to consolidate the work done within the Preservation Planning, Preservation Characterisation and Testbed subprojects. This working group became known as the Digital Object Properties Working Group (DOPWG) in early 2009 in order to reflect discussions about the terminology used to describe properties. Initially a Testbed initiative, towards the end of 2009 a shared PLANETS vision, model and vocabulary for digital object properties within digital preservation was formed, and a revamped DOPWG was established under the leadership of the Preservation Characterisation subproject.

The primary aims of the DOPWG in this form have been to:

- Provide a central platform and point of contact for digital object properties work within PLANETS;
- Investigate conceptual work and assess its practical application within PLANETS;
- Help to plan the work of the PLANETS-wide Ontology, both conceptually and practically, in order to officially release a new ontology file to the PLANETS software.

2. The PLANETS Approach

The following section sets out the major work undertaken within the PLANETS project in relation to digital object properties. It does not present the background to the role of digital object properties within the field of digital preservation or the concept of significant properties. For further information and discussion on these areas, please refer to companion report, *The concept of significant properties* (PLANETS deliverable PC3 – D23A).

2.1 The PLANETS Core Registry

During the course of the PLANETS project, two new (and one enhanced) versions of what is now known as the PLANETS Core Registry (PCR) have been designed, built, tested, released and populated[1]. The PCR is a technical registry that stores core records for File formats, Software, Hardware, Compression Techniques, Character Encodings and Storage Media along with associated subsidiary records and reference information. This persistent, unambiguous technical information supports characterisation, preservation planning and preservation action functions, and provides a growing source of technical reference information to the digital preservation community.

Within the PCR it is possible to describe significant properties in the form of the inherent and instance bytestream properties of a file format, and to assign unique identifiers or PUIDs to these. Inherent properties are the properties that all files of a particular format share e.g. lossiness or ease of identification. Instance properties are those properties that are specific to a particular file e.g. image height and width or number of pages.

Component type properties can also be recorded by the PCR. A component is the smallest logical entity within an archival record that can be considered for migration, and it will have a Component Manifestation Type (CMT) i.e. the physical manifestation of these components in a particular file format. These CMTs will have properties associated with them. Using the example of a TIFF image embedded in a Word document, the component would be image, the CMT would be a TIFF image and the component type properties would be any inherent properties associated with TIFF images.

2.1.1 PCR - Validation of preservation actions using significant properties

The PCR also stores information about characterisation tools, which identify and measure the properties of digital objects. In this way it supports the automatic deployment of appropriate characterisation tools, and thus the validation of preservation actions, by enabling the significant properties of source and target objects to be measured and compared, to ensure authenticity. In the case of migration preservation actions, the resultant information about the instance properties of the source and target components, and any variance in the properties after migration, can then be associated with a particular migration pathway and stored within the PCR[a].

[1] These have been developed as enhancements of the PRONOM Technical Registry developed by TNA. Originally known as the Preservation Characterisation Registry, it became known as the PCR when it was combined with the Preservation Action Registry.

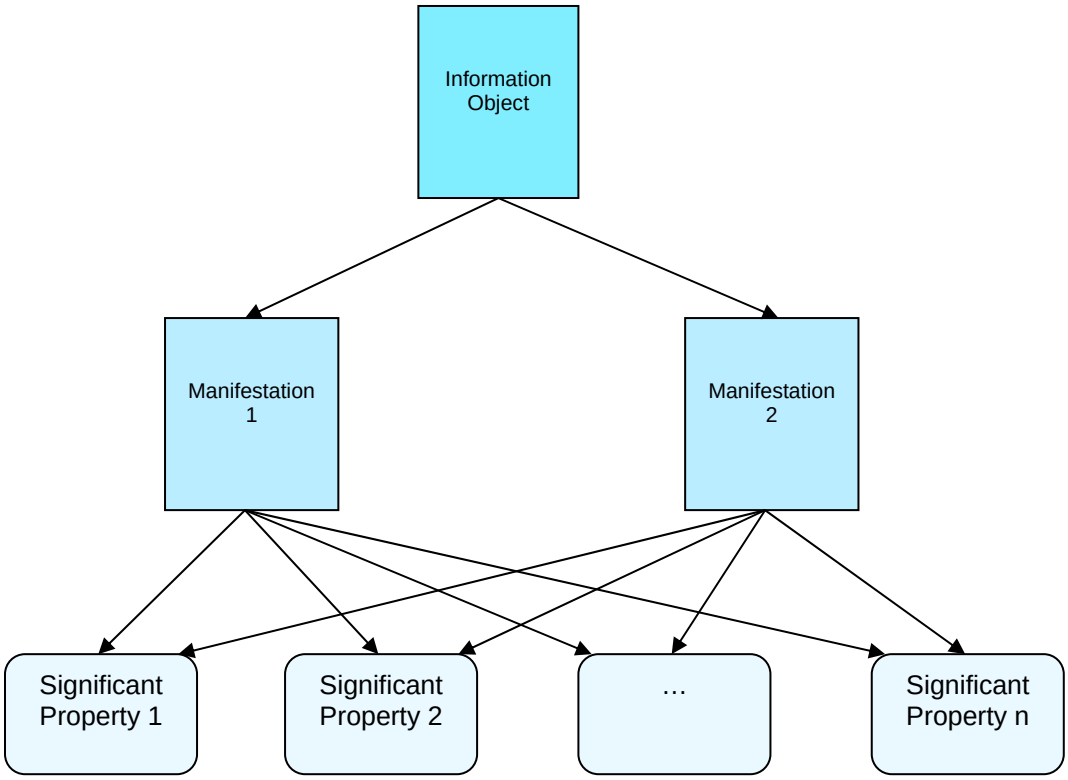

Figure 1: Validation[b]

2.1.2 PCR - Preservation planning using significant properties

The PCR has been designed to enable risk scores to be associated with properties and therefore files and formats.

The functionality of the PCR includes a risk assessment service. This was developed to provide support to Plato, the PLANETS preservation planning tool, and enables risk scores to be associated with file format and component properties, with the aim that these scores can then be used as specific node parameters within Plato-created objective trees.

These scores provide an indication of how difficult the preservation of that file or component will be. For example a file format that uses lossy compression will be harder to preserve than one with lossless compression. Each inherent property within a specific file format can be given a risk score and then the scores combined to give an overall format risk score[a].

Risk scores can also be associated with particular instances of a file, in the form of a discrete value, or a range of values, depending on the nature of the property. For example, PDF files have an inherent property that allows them to support encryption. For a particular instance of a PDF file where encryption is used, the instance risk score can reflect that this creates an increased preservation risk[c]. Alternatively, by setting value ranges for a particular property, the tolerance for a particular property can be set.

However, the work done during the course of the PLANETS project on the assignment of risk has highlighted that it is an extremely complex process. A mathematical methodology for doing this is still being developed, and therefore the risk fields within the PCR will not be populated during the course of the Project. Ultimately, whilst there may at some point be a universal concept of risk, policy will dictate organisational risk appetite, and associated risk scores will need to be decided upon accordingly[d]. It is recommended that further research be done in this area.

2.2 Preservation Characterisation

The overall goal of the Preservation Characterisation (PC) subproject was to develop methodologies, tools and services for characterising the significant properties of digital objects, in order to enable the development of preservation plans and the validation of preservation actions, and to support the needs of user communities. Characterisation can be defined as:

'*the process of determining the format-specific significant properties of an object of a given format, e.g.: "I have an object of format F; what are its salient properties?"*'[e]

The following sections highlight the major developments in the work done throughout the course of the PLANETS project to enable this, in addition to the work done within the PCR (as mentioned in the section above)

2.2.1 The XCL tools and format properties

The characterisation of digital objects is essential for both informing the development of preservation plans and validating the results of preservation actions. In the latter case, it is vital that we are able to compare the properties of a source file format with the properties of a target file format, in order to assess whether they contain the same information and consequently, whether a preservation action has been successful.

Within the PC subproject, work has been carried out to develop Extensible Characterisation Languages (XCL), and associated comparator and extractor tools. These are intended to support the automatic validation and evaluation of migrated objects as part of characterisation services provided within the PLANETS Interoperability Framework. The following section sets out this work in more detail.

Previous work and summary of pre-existing extractors

Within the PLANETS project, different strands of work have focussed on investigating the state of the art for property extraction and related topics. A summary of those efforts is presented here. The following table has been adapted from the Planets wiki[2], and updated as appropriate.

Name Tool/Service	Creator	Purpose
DROID	TNA http://sourceforge.net/projects/droid/	Identification DROID 4.0 provides automatic file format identification for over 550 formats using a combination of internal and external signatures provided by the PRONOM registry. It can perform batch identification and be operated via a GUI or Command Line Interface. Open source. DROID 5.0, provides enhanced functionality and will be available in autumn 2010.
JHOVE	Harvard University http://hul.harvard.edu/jhove/index.html	Identification/Validation/Property extraction JHOVE can validate and extract metadata from the following formats - AIFF, GIF, HTML, JPEG, JPEG2000, PDF, TIFF, WAVE, XML. It can be operated via a GUI, command line interface or Java API. Open source. JHOVE2, in its alpha release, was made available in March 2010.

[2] The table can be viewed by Planets members at http://www.planets-project.eu/private/pages/wiki/index.php/Existing_characterisation_tools

NLNZ Metadata Extractor	National Library of New Zealand http://meta-extractor.sourceforge.net/	Property extraction The Metadata Extract Tool includes a number of 'adapters' that extract metadata from specific file types. Extractors are currently provided for: • Images: BMP, GIF, JPEG and TIFF. • Office documents: MS Word (version 2, 6), Word Perfect, Open Office (version 1), MS Works, MS Excel, MS PowerPoint, and PDF. • Audio and Video: WAV and MP3. • Markup languages: HTML and XML. If a file type is unknown the tool applies a generic adapter, which extracts data that the host system 'knows' about any given file (such as size, filename, and date created). It can be operated via a GUI, command line interface or Java API. Open source.
CFX - Compound File Explorer 1.6.40	CoCo Systems Ltd http://www.coco.co.uk/developers/CFX.html	Property extraction CFX is used to explore the structure of OLE2 Compound Documents. It can be operated via a GUI. Commercial.
TrID 2.02	Marco Pontello http://mark0.net/soft-trid-e.html	Identification TrID can identify over 2,300 formats using a library of binary signatures. It can be operated via a GUI or command line interface. TrIDScan can be used to automatically generate new signatures from a sample sile set. Free for private/non-commercial use.
Metadata Miner Catalogue Pro 4.2.26	Soft Experience http://metadataminer.com/	Property Extraction Metadata Miner can extract properties from the following formats - MS Office, Visio, OpenOffice, Star Office, PDF, TIFF, JPEG, XMP. It can be operated via a GUI or command line interface. Results can be output in HTML, XML, CSV and MS Word format, and XSLT transformations are available to RDF and XSL-FO. Commercial
Jakarta POI 3.6	Apache Software Foundation http://poi.apache.org	Validation/Property Extraction The Jakarta POI project is developing a Java API to access Microsoft OLE2 Compound Document Format files, especially Excel and Word. Open source.
GSpot 2.70	Steven Greenberg http://www.free-codecs.com/download/GSpot.htm	Validation/Property Extraction Provides information on the codecs required to view an AVI movie, and also gives alerts on other possible problems. There are a range of similar tools available (see bottom of referenced web page)
Beagle 0.3.9	http://beagle-project.org	Property Extraction Beagle is a desktop search engine for linux. Beagle comes with a helper application called beagle-extract-content. beagle-extract-content is able to

		extract some fundamental properties of various file formats.
Libferris 1.2.7	http://www.libferris.com/	Property Extraction Libferris is a virtual semantic file system which is able to extract and display file information.
Statistics New Zealand Prototype PREMIS Creation Tool	http://pigpen.lib.uchicago.edu:8888/pigpen/40 (requires login and password; see: http://www.loc.gov/standards/premis/pigInfo.jpg)	Property Extraction This is a set of programs using XSL and VBScript that takes output from Jhove, the New Zealand Metadata Extractor, or DROID and produces PREMIS object records.
file(1)		Identification, Property Extraction This standard Unix tool uses magic bytes and other fixed identification schemes to recognize a wide variety of file formats and extract a few properties from some formats.
ExifTool	Phil Harvey http://www.sno.phy.queensu.ca/~phil/exiftool/	Validation, Metadata Extraction GUI and command-line batch tool. It supports EXIF, GPS, IPTC, XMP, GeoTIFF, ICC Profile and can format the output to e.g. txt, XML, XMP, JSON files or to std output. Open Source
ImageMagick	http://www.imagemagick.org/script/index.php	Property Extraction Command-line batch processing tool for making derived formats or image manipulation, which also supports tiff characterisation and extraction of EXIF/IPTC/XMP. Open source

Most of these tools were tested by PLANETS members in early 2008.[3] In addition, the evaluation of TIFF- and PDF-specific tools were summarised in the PLANETS report, *Evaluation report of additional tools and strategies*[f], of which we report the conclusions that are relevant in our context:

- A wide range of characterisation tools was tested using both TIFF and PDF files. The results of the tests showed that for TIFF files the ExifTool program was the best batch characterisation tool. When characterising PDF files the tests showed that a single tool cannot cover all the workflows needed and so different workflows require different tools.
- A prototype of a Comparative Quality Assurance Framework has been developed and this prototype has provided much new insight into the complexity of the problem.

The experiences and conclusions of the project lead to the following recommendations:

- To continue to test characterisation tools using file formats other than TIFF and PDF. New tests could include the following file formats Jpeg2000, common sound and video formats, and external metadata files (Geo MapInfo/ESRI, XML/XMP files).
- To finalise the development of the generic wrapping framework specialised for the wrapping of characterisation tools.
- To resume the wrapping of characterisation tools using the new generic wrapping framework.

[3] The results of these tests can be viewed by Planets members at http://www.planets-project.eu/private/pages/wiki/index.php/Testing_existing_characterisation_tools

Independently from these results, another PLANETS report, *Prototype extraction tool wrapper specification*[g], examined previous approaches to extracting characteristics from files. The report observed that, while the notion of characterising a file[4] seemed to be a rather clear concept, at least within the PLANETS project (and probably within the international community concerned with digital preservation), the notion itself was not widely discussed in the engineering and computer science communities. This is the likely reason that the number of tools that extract general characteristics is still remarkably low.

According to the report, only the New Zealand Metadata Extraction Tool (NZMET) and later JHOVE supported a clear, well-documented model of characterisation. Although they were very different in their logical approaches, both clearly envisage their tools to be used at one specific stage of the OAIS model; at data ingest.

Since the concept of 'characterisation' as part of the process of ingesting digital information into a repository was considered to be unknown outside the digital preservation community, the report concludes that in this sense there were no 'characterisation' tools beyond the preservation community. While the 'identify' utility of ImageMagick supported many more formats than any tool developed within the preservation community, its output could not easily be integrated into any generalized workflow for preservation, since the characteristics extracted were determined by the definition of the file format handled, not by a general underlying model of metadata and their semantics.

One of the main objectives of the XCL effort, therefore, was to define a clear abstract model in which extracted characteristics could be expressed, and to then convert the output of existing tools into a general language of characterisation.

XCL tools and the measurement of extractable properties[h]

Apart from the necessity of introducing an abstract model for the expression of format properties, the XCL tools have also been developed with the following use case in mind: the detection of deviations between two files stored in different formats, especially on a large scale.

This is illustrated by the following figure:

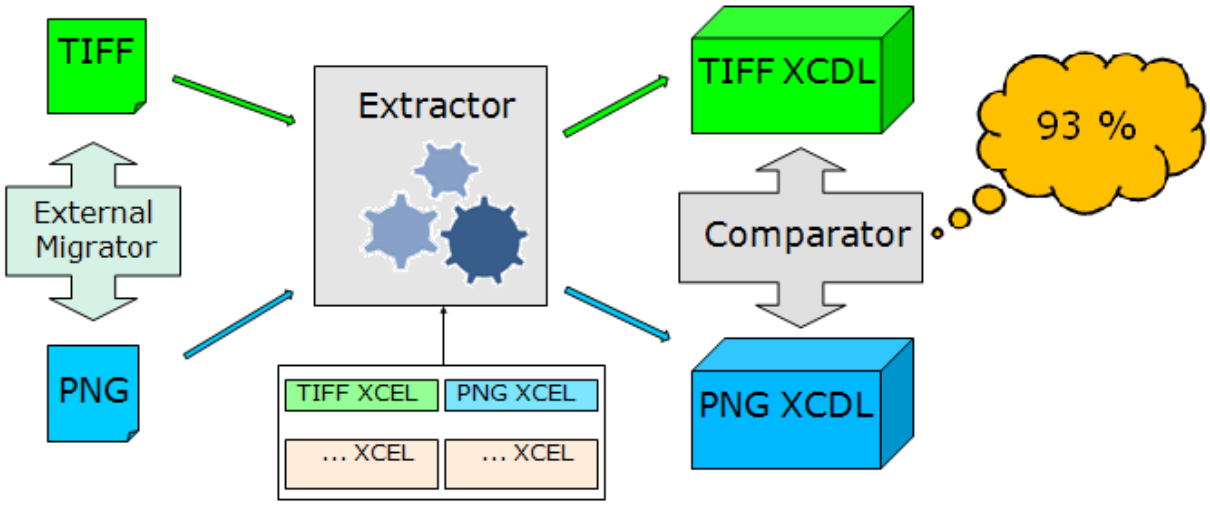

Figure 2: The XCL evaluation framework with Extractor and Comparator tools (adapted from Kurz, 2007, p. 13)[h]

The diagram shows, in three steps, the process of migration and evaluation of the data conversion through the use of the Extractor tool, the XCEL and XCDL languages, and the Comparator tool, all of which have been developed as part of the PC subproject.

[4] The deliverable states that the intuitive meaning of characterising is *"extracting from a file the small subset of data which is important for all processes necessary to preserve the complete content of that file for long periods"*

- Step 1: In this example, an image file stored in the TIFF format is migrated, into the PNG format. Now we have two image files in two different formats.

- Step 2: To enable a later comparison of the two files, the characteristics of the files need to be determined, which is done by the Extractor software. Using the eXtensible Characterisation Extraction Language (XCEL), the Extractor extracts content and technical properties such as format, width, height, size and the number of colours, from each of the files. The characteristics of each file object are saved in an eXtensible Characterisation Description Language (XCDL) document. Since XCDL handles the significant properties in a format-neutral manner, the relevant description of each image object is now available independently of the source format. Two XCDL descriptions now exist, one for the TIFF format and one for the PNG format.

- Step 3: The success of the migration is assessed by the Comparator software, which compares the XCDL descriptions of the property values and thus the characteristics of the original TIFF file against the newly created PNG file. If the characteristics did not match, the Comparator will provide a differentiated analysis of which properties have not migrated properly. In the figure above, the migration was 93 % effective. The user will have to decide whether the result of the migration is acceptable in this form or not.

The XCEL and Extractor are thus used to extract the relevant characteristics from the original and the migrated files, and the XCDL and Comparator are used to compare these properties to evaluate the success of the migration.

From what is described above, an XCDL definition of a digital object property can be extracted.

'An XCDL property is the description of the value an abstract characteristic of some data takes within a specific file. We could also say that properties describe which values of a potential characteristic shall be applied to which part of the uninterpreted data.'

The PC-specific ontology

From what has been discussed, it is clear that a mechanism to map the individual terminology of different file formats to a format-independent vocabulary is necessary to compare results. The PC subproject developed a specific ontology to solve this problem.

The XCL Ontology is the tool used to map the "surface terminology" of individual file formats (that is, the set of terms used by that specific format to describe the characteristics of the data contained in a given file) to a format-independent vocabulary. The current version of the Ontology is available at: http://planetarium.hki.uni-koeln.de/public/XCL/ontology/XCLOntology.owl.

The XCL Ontology is manually built to support the development of the XCL programs: the names, units, and data types are drawn from file format specifications for each format that the XCL suite supports. Through the development process of the Extractor and Comparator, the Ontology was used to support the automated generation of name Libraries. A web service that maps all properties stored for a certain file formats in the ontology to the so-called XCL-properties is available at http://planetarium.hki.uni-koeln.de/testbed/ontology_extract/.

These XCL-properties denote the unified property names for the output in the XCDL- and XCEL-languages. It is clear that a general controlled vocabulary of file format properties has advantages for other applications within as well as outside of the PLANETS project.[5]

2.2.2 PLANETS classification scheme for representation information networks[i]

Based on recommendations made in the White Paper on Representation Information[c], DOPWG member Henk Matthezing conducted a comparison of several representation information registries (RIRs), and put together the basis for a PLANETS representation information classification scheme. The RIRs used for the comparison were chosen because of their significance within the digital preservation field and were:

[5] For a more detailed description etc. see the Planets deliverable *eXtensible Characterisation Language Suite report*[h]

- OAIS;
- The Global Digital Format Registry (GDFR);
- The National Library of the Netherlands'/Koninklijke Bibliotheek (KB-NL) Preservation Manager;
- The web-based registry of the Library of Congress;
- PRONOM; and
- The PC-specific ontology (see section 2.2.1 above)[6].

The findings of his research revealed that PRONOM had the widest coverage for an RIR, including formats, technical components and significant properties,[7] but that it only contained top-level conceptual information. It was suggested that a combination of the GDFR, the top-level structure suggested in the White Paper, the file format concept classes from the PC-specific ontology, the Library of Congress registry and the KB-NL concept classes together provided the best multi-level structure. An integrated classification scheme was outlined and transformed into an example OWL ontology.

2.3 The Testbed

The aim of the Testbed subproject is to provide a dedicated research environment in order to allow PLANETS partners to execute, automatically evaluate and reproduce experiments, and to make available both these experiments and their documentation for long-term reference. The Testbed can be downloaded as part of the PLANETS software suite and the intention is that it will continue to be offered as an online service beyond the project. Thus, with increasing usage, it will continue to provide both information on the behaviour of preservation tools and make these tools available through the Testbed.

2.3.1 The use of properties in the Testbed to date and the coupling of machine-readable and human-observable properties

The usage of digital object properties in the PLANETS Testbed has changed considerably since the concept of properties was introduced.

Initially the Testbed produced a static list of properties associated to digital object types, a heritage from the previous Dutch Testbed project. The aim was that evaluation criteria could be construed from this in order to determine the correct characterisation, migration and rendering of a digital object's properties[i].

However, in the third year of PLANETS the XCL workflow became available, consisting of the Extractor, the Ontology and the Comparator (see section 2.2.1 above). It was deemed that the automation and the dynamics this workflow provided would bring many advantages over working with a static list. To this end work began on an extended version of the XCL ontology, from now on referred to as the Testbed ontology, which was made directly available from the front end of the Testbed interface. The Testbed ontology focused on making observational properties, such as "presence of artifacts in a converted image", available to the end-user in order to facilitate human evaluation[8].

This approach revealed itself to be unviable, given the size of the ontology to be displayed to the user. The user should not be presented with hundreds of properties for each given file type / file format. Moreover, a shift in focus towards automatically extractable properties (and therefore towards experiments based on them) meant that the issue of the representation of observational properties was no longer the first priority for the Testbed team. The end results in the final version 1.2 of Testbed are that properties from the Ontology can be automatically extracted using the Extractor. The Comparator then displays the results, which the user can then assign evaluation tags to (similar, not similar etc.).

[6] It should be noted that the PC-ontology has been developed considerably since the report was written.

[7] Although significant properties are not classed as representation information – see section 2.3 of associated report, *The concept of significant properties.* (PLANETS deliverable PC3 – D23A)

[8] See section 3.2 of the associated report, *The concept of significant properties.* (PLANETS deliverable PC3 – D23A) for discussion of observational and extractable properties.

Furthermore, a static list of observational properties will be incorporated. These will have as sources the templates of properties offered by the PLANETS preservation planning tool, Plato, and the original 'Benchmark Goals' list that was put together in the early days of the Testbed. The main issue here remains how to manage user-defined properties. In a first model, Testbed users could be allowed to create and make use of their own properties, thus ensuring that the Testbed can still be useful to future users beyond the end of the project. In a second one, any new observational properties that are suggested would have to be approved by the DOPWG – which of course poses questions of sustainability beyond the end of the PLANETS project.

In this final release of the Testbed application, both observational and extractable properties will operate in the same way. The property has a name and a type of accepted value (Boolean, integer etc).

The Testbed team has also taken into consideration the middleware and hardware sides of property evaluation for both observational and extractable properties. Both observational and extractable properties will be recorded for a particular hardware and software environment as there is no guarantee that the same measurements would result if one or more components of this environment was changed. For example, if a person is manually recording whether certain parts of an image are visible to the naked eye, this will depend on the monitor the person is using. Whether it's an LCD or a CRT may affect the 'fuzziness', as will many other considerations. And if the property is extractable, the value will still depend, not only on the tool used to make the measurement, but also on the hardware and software setup.

2.3.2 The measurement of observational properties and related theoretical issues

As stated above, it has been one of the aims of the Testbed subproject to introduce observational properties into its experiments, and to this end the Testbed ontology was developed. The Testbed ontology aims to take these and other properties into account, including service, hardware and middleware properties.

A clear distinction should be made between two concepts that are often confused when discussing this topic: observational properties and subjective properties. The Testbed subproject has always aimed to make the evaluation of the performance of digital preservation tools as scientific and objective as possible; therefore, the observational properties considered by the Testbed are those which can be assessed objectively, such as “presence of artefacts in a converted image”, “flickering”, etc. Other properties, such as “quality of colour”, though observational, introduce too large an element of subjectivity and so do not fit the context of a PLANETS experiment.

Use of results aggregation and average scores will make it much easier for a curator to evaluate the outcome of an experiment or series of experiments concerning, for instance, the average performance of a given tool on a given format. They are particularly important for making experiments on observational properties as objective as possible as any subjectivity is reduced by averaging the data from many experiments.

2.3.3 The Testbed in relation to the PCR and Plato

The Testbed could also play a role in a linked environment with the other PLANETS components PCR and Plato. In the PCR, wrapped tools are registered. Basic information about a tool, such as developer, location, version etc. will be entered into the database, with one of the most important pieces of information being the Pathway. A Pathway describes one or more preservation actions that are performed on a specific object, possibly resulting in an object of a different file format (in the case of a "tool for objects" PA tool).

One intended task of the Testbed was to aggregate the Testbed results and output these to the PCR. Although it was not possible to implement this within the course of the Planets project, this aggregation between Testbed and the PCR was specified in a draft Software Requirements document, the details of which can be found in the Planets deliverable *Planets Core Registry: Future Vision Document*. With these aggregated results the PCR could provide insight into how tools perform in various pathways. Another possible interaction would be for the preservation

planning tool PLATO (see section 2.4.2 below) to tap into the testbed information in the PCR to match that information to weighed criteria in order to present a user with the best possible tools and actions for his preservation needs.

Therefore, in this situation, a question such as "Is PDF/A suitable for long term preservation in my institution?" is a question for PLATO. In the Testbed experiments would be run on PDF/A. Related tools in the controlled hardware and software environment would be run and their behaviour evaluated against digital object properties. The objective results i.e. which would not include value judgments, would be aggregated and stored in the PCR
.
After an institution has completed an objective tree to answer the question "What is suitable for long term preservation in my institution?" PLATO would extract information from the Core Registry to give guidance. For example, it might give the answer that "a migration to PDF/A using tools X and Y is suitable for long term preservation in your institution".

The implementation of this loop with the PCR is a subject for future work. What is in place at the end of PLANETS is that Testbed can aggregate experiment outputs and make these available to Plato as an XML file.

2.4 Preservation Planning

2.4.1 Preservation Policy and Strategy Models[k l m]

The PLANETS Preservation Policy and Strategy Models workpackage developed a conceptual model in which the main concepts and requirements for digital preservation were expressed, including the modelling of technical and organisational properties. In developing the model, it was emphasised that for digital preservation to be successful, it is not solely the technical elements of digital objects that are taken into account by organisations, but also those of the environments in which the digital objects exist. To this end it is necessary to understand *'the cultural and institutional framework in which data, documents and records are created, managed and preserved'* which would include analysis of not just a digital object and its format but also the *'goals and limitations of the institution, features of its user community, and the environment in which its users access digital content'*[l].

In accordance with the model, the set of preservation guiding documents that should be considered in clarifying a stakeholder's preservation requirements may include:

- Those having any institutional scope e.g. corporate, departmental, project-related
- Those having any business focus e.g. policy, strategy, mission, process
- Those which provide a business process input to preservation planning.

These may include both written and oral representations.

Once specified, these institution-specific Preservation Guiding Requirements, which will include the constraints and tolerances that should be applied to a digital collection, can be used to inform the preservation planning process. They will help to identify risk, to identify appropriate preservation actions to mitigate risk, and to evaluate preservation actions in order to pinpoint the most suitable in a given situation. In the latter case, part of this evaluation comes from looking at the cost of losing any significant characteristics by undertaking the action.

Significance is in the eye of the Stakeholder (2009)[n]

Feeding into this conceptual model work was the paper *Significance is in the eye of the Stakeholder* (2009). The paper stemmed from the lack of clarity in terminology and meaning surrounding the concept of significant properties, and in order to clarify the situation, Dappert and Farquhar have defined a simple vocabulary for the most relevant terms:

- Entity – Anything whatsoever.
- Class – A class is a set of entities.
- Individual - Entities that are not classes are referred to as individuals.
- Property – A property is an individual that names a relationship.
- Characteristic – A property / value pair associated with an entity. The value is an entity (see figure 12 below)

Using these definitions, the property is an abstract thing that cannot on its own be preserved. It is only the characteristic, with both a property and a value, which is capable of being preserved[9].

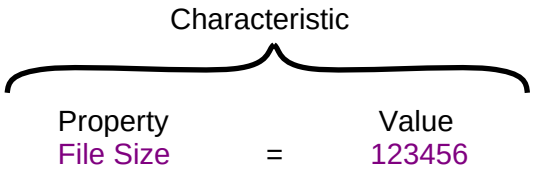

Figure 3: Dappert and Farquhar's Properties and Characteristics

According to the conceptual model proposed, any of the core classes defined i.e. Preservation Object, Preservation Action or Environment, can have properties and characteristics. Properties can also be relevant to any of the subclasses of the Preservation Object i.e. physical objects, representation objects and logical objects. The relevant stakeholders will assign significance to these characteristics according to the context in which they need to be preserved and the requirements of the stakeholders.

The paper goes on to make a number of further observations about the nature of significant characteristics[10]. Amongst these are the points that significance itself is relative i.e. a stakeholder will deem some characteristics as more significant than others; that there may be levels of tolerated deviation from the original which are acceptable; and that it will sometimes be possible and desirable to introduce and capture significant characteristics which were not found within the original digital object, but which are found within the format into which it is migrated, and which improve it. In making these observations, the paper concludes that significant characteristics are a subclass of preservation guiding requirements. The definition of significant characteristics defined by the paper, based on the observations made is:

'Requirements in a specific context, represented as constraints, expressing a combination of characteristics of preservation objects or environments that must be preserved or attained in order to ensure the continued accessibility, usability, and meaning of preservation objects, and their capacity to be accepted as evidence of what they purport to record.'

2.4.2 Preservation Planning using Plato[o] [p]

The PLANETS project takes the same approach to preservation planning as the Delos project[11] but uses a refined three-phase, 11-step workflow compared with the eight-step Utility Analysis workflow used in Delos.

As with the Utility Analysis Workflow, the 'Identify Requirements' stage of the workflow uses the structuring of an objective tree in order to clearly define the preservation goals and requirements of a specific institution for a particular collection of digital objects, and will involve identifying the associated significant properties when developing each tree. Each tree will differ depending on the individual preservation scenario, but at the highest level there will usually be four main categories of characteristics: file characteristics, record characteristics, process characteristic, and costs. The

[9] Note that the unit of measurement is also vital for the value of the property to be meaningful.

[10] See section 3.2 of the associated report, *The concept of significant properties.* (PLANETS deliverable PC3 – D23A) for discussion of the different sources of properties.

[11] As discussed in section 3.1.8 of the associated report, *The concept of significant properties.* (PLANETS deliverable PC3 – D23A).

record characteristics category, which wasn't used in the Utility Analysis Workflow, is used here to describe the technical basis of the record e.g. the context, interrelationships and metadata.

It is recommended that a workshop environment is used to gather these requirements, based on an evaluation of a representative sample of objects, and that a wide variety of different stakeholders contribute. For example this stage may bring together curators, IT administrators, consumers and other experts in the field, in order to collect a wide range of requirements. Once the tree is created, units of measurement are assigned to each leaf. It is preferable, but not always possible, for objective measurements to be used. A typical objective tree would have between 50 and several hundred objectives, arranged across a hierarchy of four to six levels.

By completing the 11 steps of the process, a ranked list of alternative preservation solutions is produced for a given preservation task, which takes into account the specific institutional requirements.

This preservation planning workflow is integrated into the PLATO Preservation Planning Tool, which can use external services to automate the process. It also adds a fourth step to the process by producing an executable preservation plan. The tool is integrated with the PLANETS interoperability framework which enables services and registries to interact with each other. In principle, this can allow PLANETS registries to be accessed for information discovery, for example:

- to provide information about the risk associated with inherent and instance properties[12];
- for preservation and characterisation services to be utilised e.g. in order to measure significant properties; and
- for a knowledge base of reusable preservation planning templates to be built in order to proactively help the planning process.

2.4.3 **Validation framework[q].**

Within the preservation planning subproject a validation framework was developed, in order to map the intellectual analysis of properties defined in the objective trees (see section above) to the technical characteristics extracted and described by the extensible characterisation languages (see section 2.2.1). The two main components of this framework are comparison metrics and an evaluation framework. These two components together support the automatic evaluation and validation of preservation actions, a process that previously has had to be done manually.

The comparison metrics are used to compute simple and aggregated measurements for low-level, technical characteristics. From the bottom-up, objects are characterised and expressed in XCDL. From this, metrics can be computed. From the top-down, preservation requirements in the form of significant properties are specified in objective trees. The evaluation framework provides a mapping between the requirements and the metrics.

The following definitions are provided:

Significant properties are defined by the user in an objective tree. They start at an intellectual level such as "object appearance" and are broken down to the criteria level, i.e. the lowest level of an objective tree where a measurement scale is specified.

Technical characteristics can be extracted from an object by an algorithm and thus a software tool.

The mapping process is done in three stages. Firstly, measurable file format characteristics and their associated comparison metrics are queried. They are then mapped to the objective tree criteria. Lastly, the comparison service is called during the evaluation of preservation actions, and a list of relevant characteristics and comparison metrics are provided, in order to assess the similarity between the original and the transformed objects.

[12] However, as stated in section 4.1.2 above, the risk assessment service will not be implemented during the Planets project.

The quality of results following a preservation action can vary considerably from tool to tool and often information is lost during the process. As seen above, part of the preservation planning process is the evaluation of planning solutions against stated requirements; in order to make this evaluation documents need to be compared before and after the preservation action takes place. The XCL languages described in section 2.2.1 above automatically support the validation and quality evaluation of preservation objects from different sources, thus allowing this comparison[13].

2.5 The integrated PLANETS-wide ontology[14]

2.5.1 The purpose of the ontology and the merging of previous effort

In the course of the PLANETS project, efforts to establish a common terminology for digital object properties among the different subprojects were only undertaken after the starting of the DOPWG activities. The aim of the merging activity was to develop an ontology that could take into account the properties of all digital object, middleware, hardware and processes relevant to both the different subprojects, and the wider digital preservation community, with the view that it could possibly be extended by future projects.

Initially the merging was done between the PC ontology (a classification containing a concept class structure for file objects, software, hardware etc) and the Testbed ontology (a file format class structure relating to observational properties). Later the effort was extended to the other property schemes developed by the PP subproject, and also to those identified by the InSPECT project. In addition, the work done in PC on the Classification scheme for representation information networks (see section 2.2.2 above), and on the properties from the PC Extractor and Comparator tools, was also taken into account.

The current version of the PLANETS-wide ontology is available at http://gforge.planets-project.eu/svn/xcltools/trunk/PLANETS_Ontology/. Log in with username: anonymous and password: empty. At http://planetarium.hki.uni-koeln.de/planets_cms/ontology/owlDoc/index.html a more user-friendly version that does not require the usage of the Protégé software is also available.

2.5.2 Benefits of ontologies and of the OWL format for property representation[h]

In principle, solutions different to the use of an ontology might have been selected for the modelling of property definitions within PLANETS: for instance, an alternative would have been using a taxonomy. However, taxonomies only model the relations between entities through inheritance, and therefore they only enable the operator to model subclass/ superclass relations between entities so as to indicate the membership of certain entities to one superclass.

Ontologies instead enable the modelling of additional relationships between entities and classes through “ObjectProperties”, these relationships can have specific names and be further separated into class hierarchies. Moreover, through the usage of reasoner software, implicit subclassing can be calculated (that is, if class A is “sameAs” class B, then all instances of A are also calculated as instances of B), and quantifiers such as “exists” and “for all” can be employed.

More specifically for the XCL suite, having an XCL-Ontology enables the software to automatically generate the names libraries (namesLibs) that are essential for the XCLsoftware. Automatically generated namesLibraries based on the Ontology would then guarantee that property names and values are valid and non-ambiguous in the XCDL- and XCEL-files.

In terms of file formats, most real-world ontologies in the area of information science are created in the Web Ontology Language (OWL). OWL is based on the RDF-language and is therefore mostly coded in XML – although it would be theoretically possible to code it in any other language which includes the required grammar. Additionally, the XML-based OWL-language makes it much easier

[13] For formats for which there is an XCEL specification

[14] See companion PLANETS report *Specification of a Planets-wide Ontology of properties for digital preservation needs.* (PLANETS deliverable PC3 – D23C) for the full ontology report.

to parse the XCL-Ontology for different purposes, and to connect it to other applications like the PLANETS Testbed.

Furthermore, OWL is a Semantic Web standard for knowledge representation. The basic idea behind the Semantic Web is to enable the searching and retrieval of information by computers based on meaning, instead of the lexical strings current search engines use. By choosing OWL for building an ontology, and using its formalised structure to relate meaningful concepts to each other, this type of ontology can be used by computers, by means of reasoning software, to help the end user navigate more easily to the information s/he needs. Publishing the PLANETS ontology on the Semantic Web would make this possible and would also enable the reuse of (parts of) the ontology by others. As there is no other ontology like the PLANETS one on the Web as yet, this reusability aspect could be of great benefit to future projects, both in the realm of digital preservation and in a wider Web perspective.

3. Recommendations for future development

Recommendations for future work have been included, where relevant, within the appropriate sections of this report, above. This chapter includes further suggestions for future development that have not been covered previously.

3.1 Further extension and refinement of the PLANETS ontology

"*A precise, finished ontology stated in a formal language is as unrealistic as a finished computer system*".[r]

Precision and clarity are the goals of the process of ontology building, but as soon as these are achieved, and there is a clear and consistent way to describe a concept, it is likely that thinking about the concept will have moved on. This seems to be inherent in the process of acquiring knowledge itself. When new knowledge within a domain is developed from a theory and its concepts, this new knowledge is very likely to alter the meaning of these concepts and the theory built upon it. This is visible in the shift of thinking from significant properties to observational and extractible properties as mentioned in section 6.1 above. Having achieved a new and more accurate way to describe properties, further thinking about them immediately clouds their meaning and their relationships with other concepts within a Knowledge Representation structure.

Knowledge in the area of digital preservation will change over time due to the fact that the technology producing digital objects is evolving continuously. Ontologies representing this domain will evolve accordingly, and this will mean that in general these ontologies will need regular maintenance. For this reason, the PLANETS ontology and the underlying ontologies on which it has been built will need revision from time to time. In addition, the work undertaken during the PLANETS project to conceptually relate file formats to their properties has not yet been completed.

Further work on ontology development is recommended in the following areas:

- Defining the boundaries of the digital preservation knowledge domain, as far as possible, in order to prevent the ontology growing in directions not directly related to digital preservation. For example, describing and classifying all hardware components should be an exercise in its own right. It can then be decided, within the boundaries of the digital preservation domain whether e.g. the classification of hardware pointing devices is relevant and should be included in the ontology.

- The PLANETS-wide ontology has been built on the various ontologies developed in the PLANETS subprojects. These sub-ontologies have been created with different visions, needs and goals and to varying degrees they have been imported, and their structure integrated, into the PLANETS-wide ontology. Further study is needed on both the

development of the sub-ontologies within their original domains, and on the level of their integration in the PLANETS-wide ontology.

- The classification structure is still incomplete, for example there is still much work to do to complete software and hardware superclasses within the digital preservation knowledge domain. This also applies to emulation and migration superclasses, which will be thinly populated initially, but will require further populations when more software packages become available. Within the Testbed ontology, classifying and relating file objects to their respective properties is also not yet finished. There are still file objects which need to be classified, and properties which need to be both described and linked to the file object classes.

- The classification structure as it is now may need revising when new scientific insights arise.

- In addition to populating the class structure, the relationships between respective properties will need to be clarified. For example software/hardware platforms are part of the process of making digital preservation objects consumable for human beings; at the same time however they are (digital) objects in their own right, and thus the process-object duality issue mentioned in section 6.1 will play a roll in describing and relating these properties.

- Finally, opportunities for the future use of the PLANETS ontology in the digital preservation domain, and its availability on the Web in order for it to be reused and adapted, need to be investigated.

3.2 Further work on a user-friendly representation of the ontology

For the representation and editing of the ontologies mentioned above, the Protégé software (http://protege.stanford.edu/) has been used. Protégé is an Open Source tool with a large and lively user and developer community. There are other commercial editing tools like TopQuadrant's TopBraid and Altova's XML editing platform, but they are relatively expensive and so were decided against.

Whilst it is clear that Protégé is possibly the only viable solution for ontologies developed in a collaborative environment at the moment, the experience of the DOPWG has shown that Protégé cannot be regarded as the most desirable option for this task, due to the steep learning curve involved in both understanding the OWL language, and the complexity and flexibility of the editor. Training and previous expertise with ontologies are required for its usage, and the interface is sometimes confusing even for experienced users. In this respect however, the other two editors mentioned do not serve as an alternative, as using them is just as complicated.

Therefore, one of the main recommendations made by DOPWG members is for a simpler interface for the editing of the ontology. Ontology editor development is a much larger domain than digital preservation and there are a number of graphical visualisation tools available. W3.org lists various tools on its W3C Semantic Web Tools page,[15] including some links to third party lists. It is recommended that these be investigated to see if there are any useful tools for the successors to the PLANETS project.

Even if there is a viewer or editor available that is simpler to use, the OWL format and its possibilities are still difficult to grasp. Further ontology development may be hindered by the lack of expertise in the field. A solution may be for non-OWL experts to develop class structures by hierarchically organising relevant concepts in a spreadsheet, and having an OWL expert or software developer develop a script for transforming this spreadsheet into the RDF/OWL language. Such a procedure has been followed manually (not involving scripts) in developing the initial PC ontology, which proved to be a lot faster than building it as an OWL ontology in Protégé, because of the large numbers of classes and individuals involved. This may be an efficient way of developing ontologies in future within the digital preservation community.

[15] See http://www.w3.org/2001/sw/wiki/Tools

3.3 Mapping of automatically extractable properties between different extractors

As described in section 2.4.3 above, work to map automatically extractable properties between different extractors was undertaken within the Preservation Planning subproject, and a basic metric and evaluation framework was developed. So far it is predominantly the practical outputs of other PLANETS work which have been connected within this common framework; more specifically, significant properties, as they are defined in objective trees, have been mapped to the technical characteristics described by XCL. Whilst within the XCL Ontology some work has been started, to look at the properties of ImageMagick and JHOVE, this is incomplete with regards to both these tools and other extractors. Therefore it is strongly recommended that future projects focus on mapping properties between external extractors.

3.4 Potential development of the PCR

A separate report has been written which sets out a future vision of possible developments for the PCR based on both those elements of the PCR which were previously specified, as part of the design for PCR 3[16], as needing further development, and further ideas for potential development and improvement. Whilst this document is not specifically related to digital object properties, some of the changes specified would impact how properties could be used within the system. Please refer to the report for further detail[s].

[16] Whilst design requirements for PCR3 were produced, it was not possible to undertake the development of the software within the scope of the Planets project.

4. Appendix - Use Case

4.1 Aims

The following narrative aims to illustrate some of the many possible interactions between a digital preservation professional and the PLANETS ontology, whilst highlighting the role of significant properties within this process. It demonstrates a potential planning and implementation scenario, written with minimal technical details, with the aim that stakeholders (for example creators of file formats, creators and curators of files, users of files, preservation policy officers, preservation plan developers, migration tool developers or characterisation tool developers), might have a common example scenario on which to focus their discussions.

4.2 User Scenario

Government institutions need to comply with legal obligations with respect to the long-term availability of their assets. Additionally, many of these institutions have started to digitise their material in order to improve access or reduce asset depreciation through heavy use. More and more specifications regulate the way that digital information has to be made available for the long term.

The National Library of Preservia, for example, hosts millions of digital documents using state-of-the-art institutional repository software. Several terabytes of data maintained in different file formats need to be managed. Amongst these are thousands of medieval documents that have been digitised. Mr. Smith is the Head of Digital Preservation and is responsible for establishing an institutional preservation strategy for the library.

Recently, Mr. Smith discovered that thousands of scanned pages of incunabula from some important medieval philosophers are stored in a file format that will not be supported by future versions of popular image viewers. According to national agreements, they have been acquired from the archive of Preservia Minor to ensure that they are safe for future use. Now, to avoid incompatibility issues and loss of information, the files need to be migrated to a new standard.

Fortunately, the National Library of Preservia participates in a large EU funded project, Galaxies, that addresses core preservation challenges and therefore Mr. Smith decides to consider some of the tools and concepts that have been developed in the course of the project so far. Within the project, a lot of effort has been invested in acquiring knowledge about file properties. This is important, because these properties are sensitive to file format migration activities. Consequently, the individual working groups of the project have put together comprehensive information about different file formats, which needs to be shared with other project partners.

Mr. Smith wants to get to know a couple of the tools that have been developed within the project to plan and implement digital preservation workflows. These tools have been crafted in different groups within the project, but share a common conceptualisation of key preservation topics. Many sub-projects within Galaxies refer to concepts and terminology that are common to preservation activities. In such circumstances, a shared vocabulary is considered to be very useful; a working group has been formed to look for a way to document the shared conceptualisations.

One option is an ontology, and this has been developed within Galaxies. Ontologies are used to support communication processes; they have been developed to help organisations find a common language and understanding of important concepts. In comparison to flat glossaries or terminology lists, ontologies can have a complex thesaurus-like structure. Thus, they not only define certain notions, but can also encode complex relationships. In this way, file properties, within the Galaxies project for example, can be much better described in relation to their surrounding preservation ecosystem.

Mr Smith wonders how such a complex structure can help people to plan and implement digital

preservation work-flows, and discovers that the ontology has been made available on-line to facilitate the searching and browsing of information about file properties. It provides information about file properties and maps them to a common identifier; this makes them comparable over a large variety of file formats. Not only are file formats and their properties mentioned, but also their relationships to other preservation topics like standards and algorithms. Mr Smith sees that similar work is going on for other disciplines e.g. in bioinformatics, where a gene ontology has been developed: http://amigo.geneontology.org/cgi-bin/amigo/go.cgi.

Mr. Smith needs to find a suitable replacement file format for the scanned incunabula images he has. This format needs to comply with certain quality requirements:

- It has to provide a data model that can carry all information without loss.
- It should be accessible by a wide variety of tools
- Its properties should be easily extractable for future preservation tasks.

For that reason, Mr. Smith has a look at the on-line version of the ontology to try to find a suitable file format. After identifying the best fit, he also finds information about the migration path from the old file format to the new one. Most of the properties seem to be easy to extract, and therefore common extractor tools should be able to handle them.

Extractor tools for evaluating the migration of file formats have also been developed within the Galaxies project. These tools are part of the eXtensible Characterisation Languages framework. The eXtensible Characterisation Extraction Language (XCEL) is already in use by the national library in combination with an extractor and comparator to access and compare significant file properties. The purpose of this language is to describe the meaning and structure of file formats in a machine-readable way. The extractor tool can use these descriptions to extract properties from files that are relevant for preservation.

The ontology supports this tool, because it maintains a list of the file properties that are considered significant for preservation during migration. In addition, for each file format, these properties have been mapped to their corresponding canonical name so that they can be compared directly (i.e. similar properties which may be named differently within different file format specifications, are given a common, recognisable name within the ontology and can therefore be compared). Since the ontology is machine-readable, by exploiting this mechanism the comparator tool can keep track of the success of the migration process.

However, Mr Smith is disappointed to find that not all of the properties he considers significant are covered by existing file property descriptions. Therefore, in collaboration with the IT group of the library, an existing XCEL file is extended, to enable the extractor to capture all of Mr Smith's significant properties. Some of the new properties cannot be directly extracted from the files but need further processing. However, since the ontology can express rich relationships, it can also provide pointers to services that perform complex mappings between file property data models.

One property, the colour palette of the outdated image file format, is itself stored in a strange format. It needs to be converted to the data model that the comparator is using. Mr. Smith discusses with the IT staff whether their tool could be enhanced to handle this special data model. They look up the colour palette property for the file format in the Ontology, and find a link to a conversion process. It points to a Web Service Definition Language (WSDL) document in a preservation interoperability framework. By processing the WSDL document, an Application Programming Interface can be constructed that implements the needed conversion process. The IT staff tell Mr Smith that with minimal effort they can enhance their migration tool to cover the needed properties. The comparator tool can now be used because all properties have the right name and data model and Mr Smith can assess whether his file migrations have been successful.

.

5. References

[a] Sinclair, P. (2009). Core Registry V3: Software Requirements Document. Planets deliverable PC3-D20.

[b] Brown, A. (2008). Characterisation in Planets. Retrieved on 8th March 2010 from http://www.dpconline.org/events/significant-properties.html

[c] Brown, A. (2008). White paper: Representation information registries. Planets ref PC3–D7. Retrieved 11th February 2010, from http://www.planets-project.eu/private/planets-ftp/docs/Deliverables/3Preservation_Characterisation_(PC)/Planets_PC3-D7_RepInformationRegistries.pdf

[d] Spencer, R. (2010). The risky business of digital preservation. The application of risk models to digital file format obsolescence. Planets deliverable PP6-D13.

[e] Retrieved on 14 April 2010 from http://hul.harvard.edu/jhove/

[f] Larsen, T., Ammitzbøll Jurik, B., Skou Hansen, T. et al (2009). Evaluation report of additional tools and strategies. Planets deliverable PC4-D12. Retrieved on 15th April 2010 from http://www.planets-project.eu/private/pages/wiki/index.php/Image:PC4-D12.doc

[g] University of Cologne (2007). Prototype extraction tool wrapper specification. Planets deliverable PC4- D3, D4 and D5. Retrieved on 18th March 2010 from (http://planetarium.hki.uni-koeln.de/planets_cms/sites/default/files/PC4D3D4D5-01.doc)

[h] University of Cologne (2009). eXtensible Characterisation Language Suite. Planets ref PC2-D12 and D13 and PC4-D7.Retrieved on 19th April 2010 from http://planetarium.hki.uni-koeln.de/planets_cms/sites/default/files/PC2D12D13PC4D7-01.pdf

[i] Hoedt, M. and Matthezing, H., (2009). Planets classification scheme for representation information networks. Planets deliverable PC3-D9.

[j] Helwig, P. (2007). Test Methods for Testbed (version 1). Planets deliverable TB3-D2. Retrieved on 22 April 2010 from http://www.planets-project.eu/private/pages/wiki/images/0/06/TB3-D2-MethodsForTesting_v1.1.pdf

[k] Dappert, A. (2009), Report on the Conceptual Aspects of Preservation, Based on Policy and Strategy Models for Libraries, Archives and Data Centres. Planets ref. PP2-D3. Retrieved on 11th February 2010 from http://www.planets-project.eu/private/planets-ftp/WP_PP/PP2/DeliverablePP2D3/PP2_D3_Conceptual_Aspects_of_Preservation.pdf

[l] Dappert, A., Ballaux, B., Mayr, M., and van Bussel, S. (2008). Report on policy and strategy models for libraries, archives and data centres. Planets deliverable PP2-D2. Retrieved on 10th February 2010 from http://www.planets-project.eu/private/planets-ftp/docs/Deliverables/1Preservation_Planning_(PP)/Planets_PP2_D2_ReportOnPolicyAndStrategyModelsM24_Ext.pdf

[m] Dappert, A. and Farquhar, A. (2009). Modelling Organizational Preservation Goals to Guide Digital Preservation. Retrieved on 1st April 2010 from http://www.ijdc.net/index.php/ijdc/article/viewFile/123/126

[n] Dappert, A., Farquhar, A. (2009). Significance is in the Eye of the Stakeholder. Retrieved 24th November 2009 from http://www.planets-project.eu/docs/papers/Dappert_Significant_Characteristics_ECDL2009.pdf

[o] Strodl, S., Becker, C., Neumayer, R., and Rauber, A. How to choose a digital preservation strategy: Evaluating a preservation planning procedure. Retrieved on 2nd April 2010 from

http://www.ifs.tuwien.ac.at/~strodl/paper/FP060-strodl.pdf

[p] Becker, C., Kulovits, H., Rauber, A., & Hofman, H. (2008). Plato: A service oriented decision support system for preservation planning. Retrieved on 2nd April 2010 from http://publik.tuwien.ac.at/files/PubDat_170832.pdf

[q] Volker Heydegger, V. and Becker, C. (2008). Specification of basic metric and evaluation framework. Planets deliverable PP5-D1. Retrieved on 9th March 2010 from http://planetarium.hki.uni-koeln.de/planets_cms/sites/default/files/Planets_PP5-D1_SpecBasicMetric_Ext.pdf

[r] Sowa, John F. (2006). A Dynamic Theory of Ontology, in: Bennet, B. & Fellbaum, C, *Formal Ontology in Information Systems*, IOS Press, Amsterdam, 2006. Retrieved on 26th April 2010 from http://www.jfsowa.com/pubs/dynonto.htm

[s] Montague, L. & van Bussel, S. (2010). Planets Core Registry: Future Vision Document. Planets deliverable PC3-D24